7

Finite-Difference Methods

In this chapter, we deal with finite-difference methods for parabolic partial differential equations, including algorithms, stability and convergence analysis, and extrapolation techniques of numerical solutions.

7.1 Finite-Difference Schemes

In this section, we will discuss the finite-difference methods for parabolic partial differential equation problems (parabolic PDE problems). Usually, a parabolic partial differential equation problem is formulated as follows:

$$\begin{cases} \dfrac{\partial u}{\partial \tau} = a(x,\tau)\dfrac{\partial^2 u}{\partial x^2} + b(x,\tau)\dfrac{\partial u}{\partial x} + c(x,\tau)u + g(x,\tau), \\ \qquad\qquad\qquad\qquad x_l \le x \le x_u, \quad 0 \le \tau \le T, \\ u(x,0) = f(x), \qquad x_l \le x \le x_u, \\ u(x_l,\tau) = f_l(\tau), \qquad 0 \le \tau \le T, \\ u(x_u,\tau) = f_u(\tau), \qquad 0 \le \tau \le T, \end{cases} \tag{7.1}$$

where $a(x,\tau) > 0$ on the domain $[x_l, x_u] \times [0,T]$ and the compatibility conditions $f(x_l) = f_l(0)$ and $f(x_u) = f_u(0)$ hold. Though sometimes, a European option problem can be approximately formulated in such a way after giving some approximate boundary condition on certain artificial boundary. However, for most of the European option problems, the problems are in or can be transformed into the following degenerate parabolic partial differential equation problem:

$$\begin{cases} \dfrac{\partial u}{\partial \tau} = a(x,\tau)\dfrac{\partial^2 u}{\partial x^2} + b(x,\tau)\dfrac{\partial u}{\partial x} + c(x,\tau)u + g(x,\tau), \\ \qquad\qquad\qquad\qquad x_l \le x \le x_u, \quad 0 \le \tau \le T, \\ u(x,0) = f(x), \quad x_l \le x \le x_u, \end{cases} \tag{7.2}$$

where $a(x,\tau) \ge 0$ on the domain $[x_l, x_u] \times [0,T]$,

Y.-l. Zhu et al., *Derivative Securities and Difference Methods*, Springer Finance, DOI 10.1007/978-1-4614-7306-0_7, © Springer Science+Business Media New York 2013

$$\begin{cases} b(x_l,\tau) - \dfrac{\partial a}{\partial x}(x_l,\tau) \ge 0, \quad 0 \le \tau \le T, \\ a\,(x_l,\tau) = 0, \quad 0 \le \tau \le T, \end{cases} \tag{7.3}$$

and

$$\begin{cases} b(x_u,\tau) - \dfrac{\partial a}{\partial x}(x_u,\tau) \le 0, \quad 0 \le \tau \le T, \\ a\,(x_u,\tau) = 0, \quad 0 \le \tau \le T. \end{cases} \tag{7.4}$$

For example, the prices of vanilla European call/put options are solutions of the problem

$$\begin{cases} \dfrac{\partial V}{\partial t} + \dfrac{1}{2}\sigma^2(S)S^2\dfrac{\partial^2 V}{\partial S^2} + (r - D_0)S\dfrac{\partial V}{\partial S} - rV = 0\ ,\ 0 \le S, \ \ 0 \le t \le T, \\ V(S,t) = \max(\pm(S-E),0), \ \ 0 \le S. \end{cases}$$

Through the transformation

$$\begin{cases} \xi = \dfrac{S}{S+E}, \\ \tau = T - t, \\ V(S,t) = (S+E)\overline{V}(\xi,\tau), \end{cases}$$

the problem is converted into

$$\begin{cases} \dfrac{\partial \overline{V}}{\partial \tau} = \dfrac{1}{2}\bar{\sigma}^2(\xi)\xi^2(1-\xi)^2\dfrac{\partial^2 \overline{V}}{\partial \xi^2} + (r - D_0)\xi(1-\xi)\dfrac{\partial \overline{V}}{\partial \xi} - [r(1-\xi) + D_0\xi]\overline{V}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0 \le \xi \le 1, \quad 0 \le \tau \le T, \\ \overline{V}(\xi,0) = \max(\pm(2\xi - 1),0), \qquad\qquad\qquad 0 \le \xi \le 1, \end{cases}$$

where $\bar{\sigma}(\xi) = \sigma(E\xi/(1-\xi))$. (For details, see Sect. 2.2.5.) Clearly, this problem is in the form (7.2). Moreover, if a stochastic model

$$dS = udt + wdX$$

is defined on $[S_l, S_u]$, and the conditions

$$\begin{cases} u\,(S_l,t) - w(S_l,t)\dfrac{\partial}{\partial S}w(S_l,t) \ge 0, \\ w\,(S_l,t) = 0 \end{cases}$$

and

$$\begin{cases} u\,(S_u,t) - w(S_u,t)\dfrac{\partial}{\partial S}w(S_u,t) \le 0, \\ w\,(S_u,t) = 0 \end{cases}$$

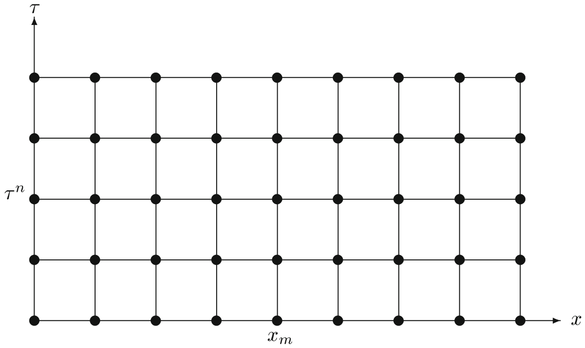

Fig. 7.1. A mesh for finite-difference methods

hold, then prices of European-style derivatives on this random variable also are solutions of the problem (7.2). (For details, see Sect. 2.4.)

To find an approximate solution of a partial differential equation problem by finite-difference methods, we first divide the domain $[x_l, x_u] \times [0, T]$ into small subdomains using lines $x_m = x_l + m\Delta x$ and $\tau^n = n\Delta\tau$, where $\Delta x = (x_u - x_l)/M$, $\Delta\tau = T/N$ and M, N are positive integers. These lines form a grid, and these points (x_m, τ^n) are called grid points (see Fig. 7.1). We want to find the approximate values of the solution on these grid points.

Let us look at the problem (7.2). First consider the case[1]

$$b(x_l, \tau) = 0, \quad 0 \le \tau \le T$$

and

$$b(x_u, \tau) = 0, \quad 0 \le \tau \le T.$$

In this case, the partial differential equation in the problem (7.2) degenerates into an ordinary differential equation at each boundary, and the degenerate parabolic problem (7.2) can be discretized in the following way.

Using forward difference for $\dfrac{\partial u}{\partial \tau}(x_m, \tau^n)$, second-order central difference for $\dfrac{\partial u}{\partial x}(x_m, \tau^n)$ and $\dfrac{\partial^2 u}{\partial x^2}(x_m, \tau^n)$ in the problem (7.2) at the point (x_m, τ^n), we have

[1] Because $a(x, \tau) \ge 0$ on $[x_l, x_u]$ and $a(x_l, \tau) = a(x_u, \tau) = 0$, we have $\dfrac{\partial a}{\partial x}(x_l, \tau) \ge 0$ and $\dfrac{\partial a}{\partial x}(x_u, \tau) \le 0$. Thus the inequality conditions in the conditions (7.3) and (7.4) can be rewritten as $b(x_l, \tau) \ge \dfrac{\partial a}{\partial x}(x_l, \tau) \ge 0$ and $b(x_u, \tau) \le \dfrac{\partial a}{\partial x}(x_u, \tau) \le 0$. Consequently, the two conditions below imply $\dfrac{\partial a}{\partial x}(x_l, \tau) = \dfrac{\partial a}{\partial x}(x_u, \tau) = 0$.

$$\begin{aligned}&\frac{u(x_m,\tau^{n+1})-u(x_m,\tau^n)}{\Delta\tau}-\frac{\Delta\tau}{2}\frac{\partial^2 u}{\partial\tau^2}(x_m,\eta)\\
&=a_m^n\left[\frac{u(x_{m+1},\tau^n)-2u(x_m,\tau^n)+u(x_{m-1},\tau^n)}{\Delta x^2}-\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi,\tau^n)\right]\\
&\quad+b_m^n\left[\frac{u(x_{m+1},\tau^n)-u(x_{m-1},\tau^n)}{2\Delta x}-\frac{\Delta x^2}{6}\frac{\partial^3 u}{\partial x^3}(\bar{\xi},\tau^n)\right]\\
&\quad+c_m^n u(x_m,\tau^n)+g_m^n,\end{aligned}$$

where

$$\eta\in(\tau^n,\tau^{n+1}),\quad \xi\in(x_{m-1},x_{m+1}),\quad \bar{\xi}\in(x_{m-1},x_{m+1}),$$

and a_m^n, b_m^n, c_m^n, and g_m^n denote $a(x_m,\tau^n), b(x_m,\tau^n), c(x_m,\tau^n)$, and $g(x_m,\tau^n)$, respectively. Dropping the term $-\dfrac{\Delta\tau}{2}\dfrac{\partial^2 u}{\partial\tau^2}(x_m,\eta)$ from the left-hand side and the two terms $-a_m^n\dfrac{\Delta x^2}{12}\dfrac{\partial^4 u}{\partial x^4}(\xi,\tau^n)$ and $-b_m^n\dfrac{\Delta x^2}{6}\dfrac{\partial^3 u}{\partial x^3}(\bar{\xi},\tau^n)$ from the right-hand side, and denoting the approximate solution of $u(x_m,\tau^n)$ by u_m^n, we obtain the following approximation to the partial differential equation in the problem (7.2):

$$\frac{u_m^{n+1}-u_m^n}{\Delta\tau}=a_m^n\frac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{\Delta x^2}+b_m^n\frac{u_{m+1}^n-u_{m-1}^n}{2\Delta x}+c_m^n u_m^n+g_m^n,$$
$$m=0,1,\cdots,M,\quad n=0,1,\cdots,N-1.$$

From the initial condition in problem (7.2), we have $u_m^0=f(x_m)$, $m=0,1,\cdots,M$. Therefore, the degenerate parabolic problem (7.2) can be discretized by

$$\begin{cases}u_m^{n+1}=\left(\dfrac{a_m^n\Delta\tau}{\Delta x^2}+\dfrac{b_m^n\Delta\tau}{2\Delta x}\right)u_{m+1}^n+\left(1-2\dfrac{a_m^n\Delta\tau}{\Delta x^2}+c_m^n\Delta\tau\right)u_m^n\\
\qquad\qquad+\left(\dfrac{a_m^n\Delta\tau}{\Delta x^2}-\dfrac{b_m^n\Delta\tau}{2\Delta x}\right)u_{m-1}^n+g_m^n\Delta\tau,\\
\qquad\qquad m=0,1,\cdots,M,\quad n=0,1,\cdots,N-1,\\
u_m^0=f(x_m),\quad m=0,1,\cdots,M.\end{cases}\tag{7.5}$$

Here, we need to point out that because we discretize ordinary differential equations at the boundaries, only u_0^n appears in the equation for $m=0$ and only u_M^n for $m=M$. That is, because $a_0^n=b_0^n=a_M^n=b_M^n=0$, u_{-1}^n and u_{M+1}^n actually do not appear in the equations above.

When u_m^n, $m=0,1,\cdots,M$ are known, we can find u_m^{n+1}, $m=0,1,\cdots,M$ by difference scheme (7.5). Because u_m^0, $m=0,1,\cdots,M$ are given in the scheme (7.5), this procedure can be done for $n=0,1,\cdots,N-1$ successively, and the approximate solution on all the grid points can be obtained. This method is called an **explicit finite-difference method**. This is because when u_m^n has been obtained, one equation involves only one unknown, so the unknown u_m^{n+1} can be computed from u_{m-1}^n, u_m^n and u_{m+1}^n explicitly.

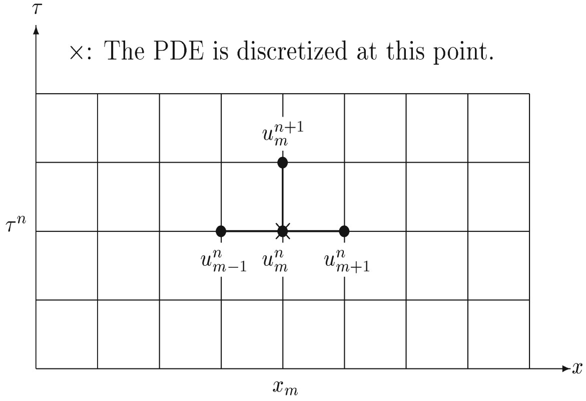

Fig. 7.2. An explicit finite-difference discretization

Figure 7.2 gives a diagram for this procedure. When we have the approximation (7.5), we have dropped the terms

$$\frac{\Delta\tau}{2}\frac{\partial^2 u}{\partial\tau^2}(x_m,\eta) - a_m^n\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi,\tau^n) - b_m^n\frac{\Delta x^2}{6}\frac{\partial^3 u}{\partial x^3}(\bar{\xi},\tau^n)$$

from the equations. These terms as a whole are called the **truncation error** for scheme (7.5). Because the truncation error can be rewritten as $O(\Delta x^2, \Delta\tau)$, we say that for scheme (7.5), the truncation error is second order in Δx and first order in $\Delta\tau$.

Now let us discretize the problem (7.2) at the point $(x_m, \tau^{n+1/2})$. For $\dfrac{\partial u}{\partial\tau}(x_m, \tau^{n+1/2})$, we use the central scheme. The derivative $\dfrac{\partial u}{\partial x}(x_m, \tau^{n+1/2})$ is approximated first by the average of the values at the points (x_m, τ^n) and (x_m, τ^{n+1}), and then the derivatives at these two points are discretized by the central difference. The second derivative $\dfrac{\partial^2 u}{\partial x^2}(x_m, \tau^{n+1/2})$ is dealt with

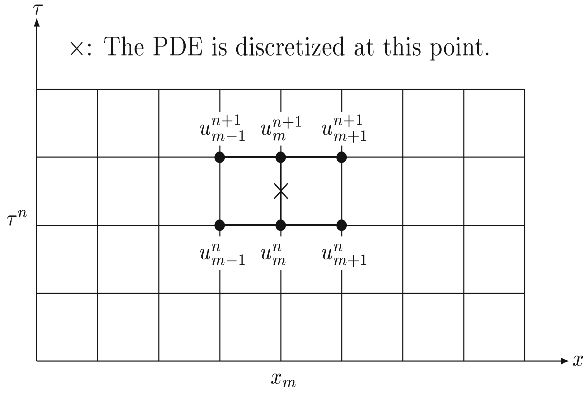

Fig. 7.3. An implicit finite-difference discretization

similarly. Using this way, the degenerate parabolic problem (7.2) can be approximated by the implicit finite-difference method:

$$\begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\Delta\tau} = \dfrac{a_m^{n+1/2}}{2}\left(\dfrac{u_{m+1}^{n+1}-2u_m^{n+1}+u_{m-1}^{n+1}}{\Delta x^2}\right. \\ \qquad\qquad \left. + \dfrac{u_{m+1}^{n}-2u_m^{n}+u_{m-1}^{n}}{\Delta x^2}\right) \\ \qquad\qquad + \dfrac{b_m^{n+1/2}}{2}\left(\dfrac{u_{m+1}^{n+1}-u_{m-1}^{n+1}}{2\Delta x}+\dfrac{u_{m+1}^{n}-u_{m-1}^{n}}{2\Delta x}\right) \\ \qquad\qquad + \dfrac{c_m^{n+1/2}}{2}(u_m^{n+1}+u_m^n)+g_m^{n+1/2}, \\ \qquad\qquad m=0,1,\cdots,M, \quad n=0,1,\cdots,N-1, \\ u_m^0 = f(x_m), \qquad m=0,1,\cdots,M. \end{cases} \tag{7.6}$$

From here, we see that each equation involves six grid points (see Fig. 7.3) and that there are three unknowns. As we know, the error of a central difference is second order. For a function, the average of the values at the points (x_m,τ^n) and (x_m,τ^{n+1}) is an approximate value at the point $(x_m,\tau^{n+1/2})$ with an error of $O(\Delta\tau^2)$ because it actually is the result obtained by the linear interpolation. Therefore, the truncation error of this scheme is $O(\Delta x^2,\Delta\tau^2)$.

Similar to the scheme (7.5), because we actually discretize ordinary differential equations at the boundaries, the equations for $m=0$ and $m=M$ can be written as

$$\frac{u_m^{n+1}-u_m^n}{\Delta\tau} = \frac{c_m^{n+1/2}}{2}(u_m^{n+1}+u_m^n)+g_m^{n+1/2},$$
$$m=0,M, \quad n=0,1,\cdots,N-1.$$

Consequently, these equations actually do not involve u_{-1}^n and u_{M+1}^n. Furthermore, the equations for $m=0$ alone can determine u_0^n, $n=1,2,\cdots,N$ from u_0^0. For u_M^n, the situation is similar. However, for u_m^n, $m\neq 0$ and M, the situation is different. We cannot determine u_m^{n+1} only from a few equations. In order to obtain u_m^{n+1}, $m=1,2,\cdots,M-1$, we have to solve a tridiagonal system of linear equations, and each of u_m^{n+1} is determined by all the u_m^n. Consequently, this method is called an **implicit finite-difference method**.

The problem (7.1) can be discretized similarly. The only difference is that the partial differential equation should not be discretized for $m=0$ and $m=M$ because the boundary conditions

$$u(x_l,\tau)=f_l(\tau)$$

and

$$u(x_u,\tau)=f_u(\tau)$$

provide the equations we need. When $a(x,\tau)$ is equal to a positive constant a, $b(x,\tau)=0$, $c(x,\tau)=0$, and $g(x,\tau)=0$, i.e., for the heat conductivity problem

$$\begin{cases} \dfrac{\partial u}{\partial \tau} = a\dfrac{\partial^2 u}{\partial x^2}, & x_l \le x \le x_u, \quad 0 \le \tau \le T, \\ u(x,0) = f(x), & x_l \le x \le x_u, \\ u(x_l,\tau) = f_l(\tau), & 0 \le \tau \le T, \\ u(x_u,\tau) = f_u(\tau), & 0 \le \tau \le T, \end{cases} \tag{7.7}$$

corresponding to the explicit scheme (7.5), (7.7) can be approximated by

$$\begin{cases} u_m^{n+1} = \alpha u_{m+1}^n + (1-2\alpha)u_m^n + \alpha u_{m-1}^n, & \\ & m = 1, 2, \cdots, M-1, \\ & n = 0, 1, \cdots, N-1, \\ u_0^{n+1} = f_l(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ u_M^{n+1} = f_u(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ u_m^0 = f(x_m), & m = 0, 1, \cdots, M, \end{cases} \tag{7.8}$$

where

$$\alpha = \frac{a\Delta\tau}{\Delta x^2}.$$

Similar to the implicit scheme (7.6), (7.7) can also be approximated by

$$\begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\Delta\tau} = \dfrac{a}{2}\left(\dfrac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2}\right. & \\ \qquad\qquad\qquad \left. + \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2}\right), & \\ \quad m = 1, 2, \cdots, M-1, & n = 0, 1, \cdots, N-1, \\ u_0^{n+1} = f_l(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ u_M^{n+1} = f_u(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ u_m^0 = f(x_m), & m = 0, 1, \cdots, M, \end{cases} \tag{7.9}$$

which is called the Crank–Nicolson scheme.

Since $u(x_l,\tau)$ and $u(x_u,\tau)$ are given, there are only $M-1$ unknowns for each time level, and the $M-1$ equations in the difference scheme (7.9) can be written together in matrix form:

$$\mathbf{A}\mathbf{u}^{n+1} = \mathbf{B}\mathbf{u}^n + \mathbf{b}^n, \tag{7.10}$$

where

$$\mathbf{A} = \begin{bmatrix} 1+\alpha & -\frac{\alpha}{2} & 0 & \cdots & 0 \\ -\frac{\alpha}{2} & 1+\alpha & -\frac{\alpha}{2} & \ddots & \vdots \\ 0 & -\frac{\alpha}{2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{\alpha}{2} \\ 0 & \cdots & 0 & -\frac{\alpha}{2} & 1+\alpha \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 1-\alpha & \frac{\alpha}{2} & 0 & \cdots & 0 \\ \frac{\alpha}{2} & 1-\alpha & \frac{\alpha}{2} & \ddots & \vdots \\ 0 & \frac{\alpha}{2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{\alpha}{2} \\ 0 & \cdots & 0 & \frac{\alpha}{2} & 1-\alpha \end{bmatrix},$$

$$\mathbf{u}^n = \begin{bmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_{M-2}^n \\ u_{M-1}^n \end{bmatrix} \quad \text{and} \quad \mathbf{b}^n = \begin{bmatrix} \frac{1}{2}\alpha u_0^n + \frac{1}{2}\alpha u_0^{n+1} \\ 0 \\ \vdots \\ 0 \\ \frac{1}{2}\alpha u_M^n + \frac{1}{2}\alpha u_M^{n+1} \end{bmatrix}.$$

Now we consider the problem (7.2) for the case

$$b(x_l, \tau) > 0, \quad 0 \le \tau \le T$$

and

$$b(x_u, \tau) < 0, \quad 0 \le \tau \le T.$$

In this case, the PDE degenerates into hyperbolic partial differential equations at the boundaries, and the first derivative there has to be discretized by a one-sided difference. For example, if in the scheme (7.5) or (7.6), we use a one-sided difference for the first derivative in the equations for $m = 0$ and $m = M$, we can have the approximation we need. We call them the modified schemes (7.5) and (7.6). However, here the way of discretizing the first derivative at $m = 0$ is different from that at $m = 1$, namely, the discretization "jumps" from $m = 0$ to $m = 1$, so from the finite-difference equation at $m = 0$ to $m = 1$, the coefficients do not satisfy the Lipschitz condition. This causes some problems when doing stability analysis. A similar situation occurs from $m = M - 1$ to $m = M$. In order to avoid the "jump," we can approximate the degenerate parabolic problem (7.2) by the explicit finite-difference method:

$$\begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\varDelta\tau} = a_m^n \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\varDelta x^2} + \varPhi_m^n + c_m^n u_m^n + g_m^n, \\ \qquad\qquad\qquad m = 0, 1, \cdots, M, \quad n = 0, 1, \cdots, N-1, \\ u_m^0 = f(x_m), \qquad m = 0, 1, \cdots, M, \end{cases} \tag{7.11}$$

where

$$\Phi_m^n = \begin{cases} b_m^n \dfrac{-u_{m+2}^n + 4u_{m+1}^n - 3u_m^n}{2\Delta x}, \text{ if } & b_m^n > 0, \\ 0, & \text{if } \quad b_m^n = 0, \\ b_m^n \dfrac{3u_m^n - 4u_{m-1}^n + u_{m-2}^n}{2\Delta x}, & \text{if } \quad b_m^n < 0 \end{cases}$$

or by the implicit finite-difference method:

$$\begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\Delta\tau} = \dfrac{a_m^{n+1/2}}{2} \left(\dfrac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2} \right. \\ \qquad\qquad\qquad\qquad \left. + \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2} \right) \\ \qquad\qquad + \Phi_m^{n+1/2} + \dfrac{c_m^{n+1/2}}{2}(u_m^{n+1} + u_m^n) + g_m^{n+1/2}, \\ \qquad\qquad m = 0, 1, \cdots, M, \quad n = 0, 1, \cdots, N-1, \\ u_m^0 = f(x_m), \qquad m = 0, 1, \cdots, M, \end{cases} \tag{7.12}$$

where

$$\Phi_m^{n+1/2} = \begin{cases} \dfrac{b_m^{n+1/2}}{2} \left(\dfrac{-u_{m+2}^{n+1} + 4u_{m+1}^{n+1} - 3u_m^{n+1}}{2\Delta x} \right. \\ \qquad \left. + \dfrac{-u_{m+2}^n + 4u_{m+1}^n - 3u_m^n}{2\Delta x} \right), \text{ if } & b_m^{n+1/2} > 0, \\ 0, & \text{if } \quad b_m^{n+1/2} = 0, \\ \dfrac{b_m^{n+1/2}}{2} \left(\dfrac{3u_m^{n+1} - 4u_{m-1}^{n+1} + u_{m-2}^{n+1}}{2\Delta x} \right. \\ \qquad \left. + \dfrac{3u_m^n - 4u_{m-1}^n + u_{m-2}^n}{2\Delta x} \right), & \text{if } \quad b_m^{n+1/2} < 0. \end{cases}$$

Scheme (7.12) usually involves eight points, among them there are four unknowns (see Fig. 7.4). However, at boundaries there are three unknowns because $a_0^{n+1/2} = a_M^{n+1/2} = 0$. When the partial differential equation is discretized in this way, the stability analysis can be done much easier. In the paper [79] by Sun, Yan, and Zhu, the stability problem of scheme (7.12) has been carefully studied. Clearly, the truncation error of the scheme (7.11) is $O(\Delta x^2, \Delta\tau)$ and that of the scheme (7.12) is $O(\Delta x^2, \Delta\tau^2)$.

Therefore, in order to find a solution, we can use either an explicit finite-difference method or an implicit finite-difference method. From the next section, we will see that for an explicit method, the step size $\Delta\tau$ must be less than a constant times Δx^2 for a stable computation. Thus, if a small Δx must be adopted in order to have satisfying results, the computation could take quite a long time. However, there is no restriction on the step size $\Delta\tau$

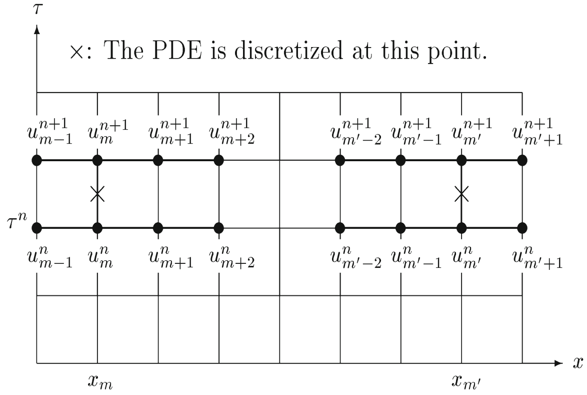

Fig. 7.4. Implicit eight-point finite-difference discretizations

for implicit finite-difference methods. This is the main advantage of implicit methods over explicit methods.

A European-style derivative could involve several random state variables. In this case, we need to discretize a multi-dimensional problem, which will be dealt with in Chaps. 8 and 10. Usually, an American-style derivative problem can be formulated as a free boundary problem. Discretization of such a problem will be discussed in Chap. 9.

7.2 Stability and Convergence Analysis

7.2.1 Stability

Stability is concerned with the propagation of errors. During the computation, truncation errors are brought into approximate solutions at each step. Also rounding errors are introduced into solutions all the time because any computer has a finite number of digits for numbers. If for a given finite-difference method, the errors are not magnified at each step in some norm, then we say that the finite-difference method is stable. There are two different norms that are often used in studying stability. Suppose

$$\mathbf{x} = (x_1, x_2, \cdots, x_{M-1})^T$$

is a vector with $M-1$ components. The L_∞ and L_2 norms of the vector $\mathbf{x}$ are defined as follows:

$$||\mathbf{x}||_{\mathrm{L}_\infty} = \max_{1 \le m \le M-1} |x_m|$$

and

$$||\mathbf{x}||_{\mathrm{L}_2} = \left(\frac{1}{M-1} \sum_{m=1}^{M-1} x_m^2 \right)^{1/2} .$$

Here, $M-1$ could be any positive integer and is allowed to go to infinity.

Stability of Explicit Finite-Difference Methods for the Heat Equation. Consider the explicit finite-difference method (7.8) for the heat conductivity problem. Suppose an initial error e_m^0 appears in computing $f(x_m)$ for $m = 1, 2, \cdots, M-1$. That is, instead of $f(x_m)$, $f(x_m) + e_m^0$ is given as the initial value. We assume that there is no error from boundary conditions, that is, $e_0^0 = e_M^0 = 0$. Let $\tilde{u}_m^n, m = 0, 1, \cdots, M, n = 0, 1, \cdots, N$, be the computed solution. We want to study how $\tilde{u}_m^n$ is affected by e_m^0. This is usually referred to as studying the stability of schemes with respect to initial values. Clearly, $\tilde{u}_m^n$ satisfies

$$\begin{cases} \tilde{u}_m^{n+1} = \alpha\tilde{u}_{m+1}^n + (1-2\alpha)\tilde{u}_m^n + \alpha\tilde{u}_{m-1}^n, \\ \qquad m = 1, 2, \cdots, M-1, & n = 0, 1, \cdots, N-1, \\ \tilde{u}_0^{n+1} = f_l(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ \tilde{u}_M^{n+1} = f_u(\tau^{n+1}), & n = 0, 1, \cdots, N-1, \\ \tilde{u}_m^0 = f(x_m) + e_m^0, & m = 0, 1, \cdots, M. \end{cases}$$

Let

$$e_m^n = \tilde{u}_m^n - u_m^n, \quad m = 0, 1, \cdots, M, \quad n = 0, 1, \cdots, N.$$

Taking the difference of the scheme (7.8) and this system, we get

$$\begin{cases} e_m^{n+1} = \alpha e_{m+1}^n + (1-2\alpha)e_m^n + \alpha e_{m-1}^n, \\ \qquad m = 1, 2, \cdots, M-1, & n = 0, 1, \cdots, N-1, \\ e_0^{n+1} = 0, & n = 0, 1, \cdots, N-1, \\ e_M^{n+1} = 0, & n = 0, 1, \cdots, N-1, \\ e_m^0 = e_m^0, & m = 0, 1, \cdots, M. \end{cases} \tag{7.13}$$

For this scheme, we can analyze its stability in two ways. First, we show that this scheme is stable in the maximum norm if $\alpha \le 1/2$. In this case, all the coefficients in the right-hand side of the finite-difference equation, α, $1-2\alpha$, α, are nonnegative, so

$$\begin{aligned} |e_m^{n+1}| &= |\alpha e_{m+1}^n + (1-2\alpha)e_m^n + \alpha e_{m-1}^n| \\ &\le \alpha|e_{m+1}^n| + (1-2\alpha)|e_m^n| + \alpha|e_{m-1}^n| \\ &\le \max_{1\le m\le M-1} |e_m^n|, \quad m = 1, 2, \cdots, M-1, \end{aligned}$$

or

$$\max_{1\le m\le M-1} |e_m^{n+1}| \le \max_{1\le m\le M-1} |e_m^n|,$$

where we have used the fact $e_0^n = e_M^n = 0$, $n = 0, 1, \cdots, N$. This is true for any n. Therefore,

$$\max_{1\le m\le M-1} |e_m^n| \le \max_{1\le m\le M-1} |e_m^0|$$

or

$$||\mathbf{e}^n||_{\mathrm{L}_\infty} \le ||\mathbf{e}^0||_{\mathrm{L}_\infty}.$$

Consequently, the difference scheme (7.8) is stable with respect to initial value in the maximum norm. This method of analyzing stability is very simple. Unfortunately, it seems that this method works only for explicit schemes with positive coefficients on the right-hand side.

Now let us study the stability of scheme (7.8) in another way. Set

$$\mathbf{A}_1 = \begin{bmatrix} 1-2\alpha & \alpha & 0 & \cdots & 0 \\ \alpha & 1-2\alpha & \alpha & \ddots & \vdots \\ 0 & \alpha & 1-2\alpha & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & \alpha & 1-2\alpha \end{bmatrix}, \quad \mathbf{e}^n = \begin{bmatrix} e_1^n \\ e_2^n \\ \vdots \\ \vdots \\ e_{M-1}^n \end{bmatrix}. \tag{7.14}$$

From the system (7.13), we see that between $\mathbf{e}^{n+1}$ and $\mathbf{e}^n$ there is the following relation:

$$\mathbf{e}^{n+1} = \mathbf{A}_1\mathbf{e}^n.$$

Suppose λ is an eigenvalue of $\mathbf{A}_1$ and $\mathbf{x} = (x_1, x_2, \cdots, x_{M-1})^T$ is an associated eigenvector, i.e., we assume that λ and $\mathbf{x}$ satisfy the equation

$$\mathbf{A}_1\mathbf{x} = \lambda\mathbf{x}.$$

Now let us find $M-1$ linearly independent eigenvectors of $\mathbf{A}_1$ and their associated eigenvalues. Define

$$x_0 = x_M = 0.$$

Then the equation above can be rewritten as

$$\alpha x_{m-1} + (1-2\alpha)x_m + \alpha x_{m+1} = \lambda x_m, \quad 1 \le m \le M-1, \tag{7.15}$$

or

$$\alpha x_{m-1} + (1-2\alpha-\lambda)x_m + \alpha x_{m+1} = 0, \quad 1 \le m \le M-1. \tag{7.16}$$

For the system (7.16) with arbitrary x_0 and x_M, let us try to find a solution in the form

$$x_m = \mu^m, \quad 0 \le m \le M. \tag{7.17}$$

Substituting it into system (7.16), we have

$$\left[\alpha + (1-2\alpha-\lambda)\mu + \alpha\mu^2\right]\mu^{m-1} = 0, \quad 1 \le m \le M-1,$$

which can be reduced to one equation:

$$\alpha\mu^2 + (1-2\alpha-\lambda)\mu + \alpha = 0. \tag{7.18}$$

Denote the two roots of Eq. (7.18) by μ_1 and μ_2. It is clear that μ_1 and μ_2 should satisfy the following conditions:

$$\mu_1 + \mu_2 = -\frac{1}{\alpha}(1 - 2\alpha - \lambda), \qquad \mu_1 \mu_2 = 1.$$

Case one: $\mu_1 = \mu_2 = \mu_*$. In this case,

$$x_m = m\mu_*^m, \quad 0 \le m \le M,$$

also is a solution of the system (7.16). Substituting it into system (7.16) yields

$$\begin{aligned} &\alpha(m-1)\mu_*^{m-1} + (1 - 2\alpha - \lambda)m\mu_*^m + \alpha(m+1)\mu_*^{m+1} \\ &= -\alpha\mu_*^{m-1} + \alpha\mu_*^{m+1} = \alpha\mu_*^{m-1}(\mu_*^2 - 1) = 0, \quad 1 \le m \le M-1, \end{aligned}$$

because of $\mu_1\mu_2 = \mu_*^2 = 1$, so it is true that $x_m = m\mu_*^m, \quad 0 \le m \le M$, is another solution of the system (7.16) besides the solution (7.17) with $\mu = \mu_*$. Thus for any c_1 and c_2,

$$x_m = (\, c_1 + c_2 m \,)\mu_*^m, \quad 0 \le m \le M,$$

should be a solution of the system (7.16). It follows from $x_0 = x_{\scriptscriptstyle M} = 0$ that $c_1 = c_2 = 0$. Consequently, $x_m \equiv 0$, $1 \le m \le M-1$, which contradicts that $\mathbf{x} = (x_1, x_2, \cdots, x_{\scriptscriptstyle M-1})^T$ is an eigenvector.

Case two: $\mu_1 \ne \mu_2$. In this case for any c_1 and c_2,

$$x_m = c_1\mu_1^m + c_2\mu_2^m, \quad 0 \le m \le M,$$

should be a solution of the system (7.16). It follows from $x_0 = x_{\scriptscriptstyle M} = 0$ that

$$c_1 + c_2 = 0, \quad c_1\mu_1^M + c_2\mu_2^M = 0.$$

From these two relations we can obtain

$$\left(\frac{\mu_1}{\mu_2}\right)^M = -\frac{c_2}{c_1} = 1 = \mathrm{e}^{\mathrm{i}2k\pi}, \quad k \text{ being any integer.}$$

Consequently,

$$\frac{\mu_1}{\mu_2} = \mathrm{e}^{\mathrm{i}2\omega_k}, \quad \omega_k = \frac{k\pi}{M}, \quad k \text{ being any integer.}$$

It is clear that $k = k^*$ and $k = k^* + M$ give the same solution. Thus we need to set $k = 0, 1, \cdots, M-1$ only. For $k = 0$, we have $\mu_1 = \mu_2$. As we have pointed out, in this case we could not find any eigenvector. For $k = 1, 2, \cdots,$ or $M-1$, we have $\dfrac{\mu_1}{\mu_2} = \mathrm{e}^{\mathrm{i}2\omega_k}$. Combining this relation with $\mu_1\mu_2 = 1$ yields

$$\mu_1^{(k)} = \mathrm{e}^{\mathrm{i}\omega_k}, \quad \mu_2^{(k)} = \mathrm{e}^{-\mathrm{i}\omega_k}.$$

For such a k, taking $c_1 = \frac{1}{2}$ and $c_2 = -\frac{1}{2}$, we have the following eigenvector

$$\mathbf{x}_{\omega_k} = \begin{bmatrix} \frac{1}{2}\mathrm{e}^{\mathrm{i}\omega_k} - \frac{1}{2}\mathrm{e}^{-\mathrm{i}\omega_k} \\ \frac{1}{2}\mathrm{e}^{\mathrm{i}2\omega_k} - \frac{1}{2}\mathrm{e}^{-\mathrm{i}2\omega_k} \\ \vdots \\ \vdots \\ \frac{1}{2}\mathrm{e}^{\mathrm{i}(M-1)\omega_k} - \frac{1}{2}\mathrm{e}^{-\mathrm{i}(M-1)\omega_k} \end{bmatrix} = \begin{bmatrix} \sin\omega_k \\ \sin 2\omega_k \\ \vdots \\ \vdots \\ \sin(M-1)\omega_k \end{bmatrix}. \tag{7.19}$$

The corresponding eigenvalue λ_{ω_k} satisfies system (7.15), i.e.,

$$\begin{aligned} \lambda_{\omega_k} &= \frac{\alpha\sin(m-1)\omega_k + (1-2\alpha)\sin m\omega_k + \alpha\sin(m+1)\omega_k}{\sin m\omega_k} \\ &= \frac{\alpha\sin m\omega_k\cos\omega_k + (1-2\alpha)\sin m\omega_k + \alpha\sin m\omega_k\cos\omega_k}{\sin m\omega_k} \\ &= 1-2\alpha+2\alpha\cos\omega_k = 1-4\alpha\sin^2(\omega_k/2). \end{aligned}$$

Here $k = 1, 2, \cdots, M-1$, i.e., we have found $M-1$ eigenvalues of $\mathbf{A}_1$ and their associated eigenvectors. Because λ_{ω_k}, $k = 1, 2, \cdots, M-1$, are distinct eigenvalues of the symmetric matrix $\mathbf{A}_1$, the $M-1$ associated eigenvectors, $\mathbf{x}_{\omega_k}$, $k = 1, 2, \cdots, M-1$, are linearly independent.

As a consequence, any vector with $M-1$ components can be expressed as linear combination of $\mathbf{x}_{\omega_k}$, which means that an error $\mathbf{e}^0$ can be expressed as

$$\mathbf{e}^0 = \sum_{k=1}^{M-1} \varepsilon_{\omega_k}\mathbf{x}_{\omega_k}.$$

Substituting this expression into $\mathbf{e}^{n+1} = \mathbf{A}_1\mathbf{e}^n$, we have

$$\mathbf{e}^1 = \mathbf{A}_1\mathbf{e}^0 = \sum_{k=1}^{M-1} \varepsilon_{\omega_k}\lambda_{\omega_k}\mathbf{x}_{\omega_k}$$

and furthermore

$$\mathbf{e}^n = \sum_{k=1}^{M-1} \varepsilon_{\omega_k}\lambda_{\omega_k}^n\mathbf{x}_{\omega_k}$$

or in component form

$$e_m^n = \sum_{k=1}^{M-1} \varepsilon_{\omega_k}\lambda_{\omega_k}^n \sin m\omega_k, \quad m = 1, 2, \cdots, M-1.$$

As eigenvectors of a symmetric matrix $\mathbf{A}_1$, $\mathbf{x}_{\omega_k}$, $k = 1, 2, \cdots, M-1$ are orthogonal. Thus, from the expressions of $\mathbf{e}^0$ and $\mathbf{e}^n$ above, we have

$$||\mathbf{e}^0||_{\mathrm{L}_2} = \left(\frac{1}{M-1}\sum_{m=1}^{M-1} \varepsilon_{\omega_k}^2 ||\mathbf{x}_{\omega_k}||_{\mathrm{L}_2}^2\right)^{1/2}$$

and

$$||\mathbf{e}^n||_{\mathrm{L}_2} = \left(\frac{1}{M-1} \sum_{m=1}^{M-1} \varepsilon_{\omega_k}^2 \lambda_{\omega_k}^{2n} ||\mathbf{x}_{\omega_k}||_{\mathrm{L}_2}^2 \right)^{1/2} .$$

Consequently, we obtain

$$||\mathbf{e}^n||_{\mathrm{L}_2} \le ||\mathbf{e}^0||_{\mathrm{L}_2}$$

if all the eigenvalues of $\mathbf{A}_1$ are in $[-1,1]$. From what we have gotten the following conclusion is obtained: if

$$0 \le \alpha \le 1/2,$$

then we have the following inequality

$$-1 \le 1-4\alpha \le \lambda_{\omega_k} = 1 - 4\alpha \sin^2(\omega_k/2) \le 1, \quad k = 1, 2, \cdots, M-1,$$

which means that the computation is stable with respect to the initial value. If $\alpha > 1/2$, then when M is large enough, some of the eigenvalues of $\mathbf{A}_1$ must be less than -1. Hence, if a component of $\mathbf{e}^0$ associated with such an eigenvalue is not zero, then the corresponding component of $\mathbf{e}^n$ will be greater than the component of $\mathbf{e}^0$ and go to infinity as n goes to infinity. Because the errors are random variables, the ε_{ω_k} corresponding to such an eigenvalue λ_{ω_k} might not be zero. Thus, the computation is unstable. This can be summarized as: scheme (7.8) is stable if

$$\alpha = \frac{a\Delta\tau}{\Delta x^2} \le 1/2;$$

whereas the scheme is unstable if

$$\alpha = \frac{a\Delta\tau}{\Delta x^2} > 1/2.$$

Stability of Implicit Finite-Difference Methods for the Heat Equation. The second method used above to analyze stability can be applied to other cases, for example, implicit finite-difference methods. For an implicit finite-difference scheme, suppose $\mathbf{e}^n$ satisfies

$$\mathbf{A}\mathbf{e}^{n+1} = \mathbf{B}\mathbf{e}^n,$$

where $\mathbf{A}$ and $\mathbf{B}$ are two matrices, and $\mathbf{A}$ is invertible. Also, assume that the following relations hold:

$$\lambda_{\omega_k} \mathbf{A}\mathbf{x}_{\omega_k} = \mathbf{B}\mathbf{x}_{\omega_k}, \quad k = 1, 2, \cdots, M-1, \tag{7.20}$$

where $\mathbf{x}_{\omega_k}$, $k = 1, 2, \cdots, M-1$ are linear independent vectors. In this case, this method still works: if all the $\lambda_{\omega_k} \in [-1,1]$, then the scheme is stable; if certain λ_{ω_k} does not belong to $[-1,1]$, then the scheme is unstable. In fact, any initial error can be expressed as

$$\mathbf{e}^0 = \sum_{k=1}^{M-1} \varepsilon_{\omega_k} \mathbf{x}_{\omega_k}$$

and because of the set of relations (7.20), we have

$$\mathbf{e}^n = \sum_{k=1}^{M-1} \varepsilon_{\omega_k} \lambda_{\omega_k}^n \mathbf{x}_{\omega_k}$$

for any n. Therefore, the scheme is stable if and only if

$$|\lambda_{\omega_k}| \le 1$$

for all the ω_k.

For the Crank–Nicolson scheme (7.9), $\mathbf{A}$ and $\mathbf{B}$ are given in Sect. 7.1. As pointed out above, in order to study the stability, we need to find the solution of

$$\lambda \mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}.$$

In Problem 7, for more general equations, readers are asked to find the eigenvectors and the eigenvalues. Here we only give the result. The result is as follows. For this case, there are $M-1$ linearly independent vectors given by the expression (7.19) and the corresponding eigenvalues are

$$\begin{aligned}\lambda_{\omega_k} &= \frac{\frac{1}{2}\alpha \sin(m+1)\omega_k + (1-\alpha)\sin m\omega_k + \frac{1}{2}\alpha \sin(m-1)\omega_k}{-\frac{1}{2}\alpha \sin(m+1)\omega_k + (1+\alpha)\sin m\omega_k - \frac{1}{2}\alpha \sin(m-1)\omega_k} \\ &= \frac{(1-\alpha)\sin m\omega_k + \alpha \sin m\omega_k \cos\omega_k}{(1+\alpha)\sin m\omega_k - \alpha \sin m\omega_k \cos\omega_k} \\ &= \frac{1 - 2\alpha \sin^2 \dfrac{\omega_k}{2}}{1 + 2\alpha \sin^2 \dfrac{\omega_k}{2}}, \quad k = 1, 2, \cdots, M-1,\end{aligned}$$

where $\omega_k = k\pi/M$. Because $|\lambda_{\omega_k}| \le 1$ for any ω_k, the difference scheme (7.9) is stable in the L_2 norm.

Stability for Periodic Problems. In schemes (7.8) and (7.9), the values are given at both boundaries, and during stability analysis, we assume that there is no error at the boundaries. It is clear that this is not always the case. Consider problems satisfying periodic conditions and assume $u_m^n = u_{m+M}^n$. In this case, we only need to find u_m^n, $m = 0, 1, \cdots, M-1$ for each time level. If the coefficients of the problem are constant, then we can analyze the stability in a similar way. Let us further assume that the solution satisfies the system:

$$a_1 u_{m+1}^{n+1} + a_0 u_m^{n+1} + a_{-1} u_{m-1}^{n+1} = b_1 u_{m+1}^n + b_0 u_m^n + b_{-1} u_{m-1}^n, \; m = 0, 1, \cdots, M-1.$$

If e_m^n is the error of u_m^n, then e_m^n satisfy the same system. Thus, the system for e_m^n can be written as

$$\mathbf{A}_2\mathbf{e}^{n+1} = \mathbf{B}_2\mathbf{e}^n,$$

where we have used the conditions

$$e^n_{-1} = e^n_{M-1}, \quad e^n_M = e^n_0$$

and adopted the following notation:

$$\mathbf{A}_2 = \begin{bmatrix} a_0 & a_1 & 0 & \cdots & a_{-1} \\ a_{-1} & a_0 & a_1 & \ddots & \vdots \\ 0 & a_{-1} & a_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_1 \\ a_1 & \cdots & 0 & a_{-1} & a_0 \end{bmatrix}, \quad \mathbf{e}^n = \begin{bmatrix} e^n_0 \\ e^n_1 \\ \vdots \\ \vdots \\ e^n_{M-1} \end{bmatrix}$$

and

$$\mathbf{B}_2 = \begin{bmatrix} b_0 & b_1 & 0 & \cdots & b_{-1} \\ b_{-1} & b_0 & b_1 & \ddots & \vdots \\ 0 & b_{-1} & b_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_1 \\ b_1 & \cdots & 0 & b_{-1} & b_0 \end{bmatrix}.$$

In order to study stability, we need to find the solution of the equation

$$\lambda\mathbf{A}_2\mathbf{x} = \mathbf{B}_2\mathbf{x}.$$

This is left for readers to do as Problem 8. The result is as follows. For this equation, the eigenvectors are

$$\mathbf{x}_{\theta_k} = \begin{bmatrix} 1 \\ \mathrm{e}^{\mathrm{i}\theta_k} \\ \vdots \\ \vdots \\ \mathrm{e}^{\mathrm{i}(M-1)\theta_k} \end{bmatrix}, \quad k = 0, 1, \cdots, M-1,$$

where $\theta_k = 2k\pi/M$ and the eigenvalues are

$$\lambda_{\theta_k} = \frac{b_1\mathrm{e}^{\mathrm{i}\theta_k} + b_0 + b_{-1}\mathrm{e}^{-\mathrm{i}\theta_k}}{a_1\mathrm{e}^{\mathrm{i}\theta_k} + a_0 + a_{-1}\mathrm{e}^{-\mathrm{i}\theta_k}}, \quad k = 0, 1, \cdots, M-1.$$

By using the relations $\mathrm{e}^{-\mathrm{i}\theta_k} = \mathrm{e}^{\mathrm{i}(M-1)\theta_k}$ and $\mathrm{e}^{\mathrm{i}M\theta_k} = 1$, this result can be shown by a straightforward calculation. If $|\lambda_{\theta_k}| \le 1, \quad k = 0, 1, \cdots, M-1$, then the method is stable. If $|\lambda_{\theta_k}| > 1$ for some k, then the method is unstable.

Because M can go to infinity, θ_k indeed can be any number in the interval $[0, 2\pi]$. Therefore, if for any $\theta \in [0, 2\pi]$,

$$|\lambda_\theta| = \left| \frac{b_1 \mathrm{e}^{\mathrm{i}\theta} + b_0 + b_{-1}\mathrm{e}^{-\mathrm{i}\theta}}{a_1 \mathrm{e}^{\mathrm{i}\theta} + a_0 + a_{-1}\mathrm{e}^{-\mathrm{i}\theta}} \right| \le 1, \tag{7.21}$$

then the scheme is stable. Otherwise, the method is unstable. Such a method of analyzing stability is usually called the von Neumann method and λ_θ is called the amplification factor. This method gives a complete stability analysis for periodic initial value problems with constant coefficients. For more general case, this method can be performed in the following way. Assume

$$e_m^n = \lambda_\theta^n \mathrm{e}^{\mathrm{i}m\theta}, \tag{7.22}$$

where θ can be any real number in the interval $[0, 2\pi]$. Substituting this expression into the finite-difference equation, we can find λ_θ. If all $|\lambda_\theta| \le 1$, then the scheme is stable; if some $|\lambda_\theta| > 1$, then the scheme is unstable. For more about this method, see the book [67] by Richtmyer and Morton and many other books.

Stability Analysis in Practice. In practice, most problems have variable coefficients. Therefore, the von Neumann method does not give a complete stability analysis. However, it is still very useful. The von Neumann method can be applied in practice in the following way.

Consider the following scheme with variable coefficients:

$$\begin{aligned} &a_{1,m}^n u_{m+1}^{n+1} + a_{0,m}^n u_m^{n+1} + a_{-1,m}^n u_{m-1}^{n+1} \\ &= b_{1,m}^n u_{m+1}^n + b_{0,m}^n u_m^n + b_{-1,m}^n u_{m-1}^n, \end{aligned} \tag{7.23}$$

where for simplicity, we assume that only three points in the x direction are involved. If more points are involved, the procedure is still the same. Suppose

$$|f_{m+1}^n - f_m^n| < c\Delta x, \qquad |f_{m+1}^n - 2f_m^n + f_{m-1}^n| < c\Delta x^2,$$

and

$$|f_m^{n+1} - f_m^n| < c\Delta\tau$$

for $f = a_1, a_0, a_{-1}, b_1, b_0,$ and b_{-1}. Assume that e_m^n has the form (7.22). Substituting this expression into the finite-difference equation (7.23) yields

$$\lambda_\theta(x_m, \tau^n) = \frac{b_{1,m}^n \mathrm{e}^{\mathrm{i}(m+1)\theta} + b_{0,m}^n \mathrm{e}^{\mathrm{i}m\theta} + b_{-1,m}^n \mathrm{e}^{\mathrm{i}(m-1)\theta}}{a_{1,m}^n \mathrm{e}^{\mathrm{i}(m+1)\theta} + a_{0,m}^n \mathrm{e}^{\mathrm{i}m\theta} + a_{-1,m}^n \mathrm{e}^{\mathrm{i}(m-1)\theta}}.$$

If for the amplification factor, we have

$$|\lambda_\theta(x_m, \tau^n)| \le 1$$

for every point and the treatment of boundary conditions is reasonable, then we can expect the scheme to be stable. Clearly, the condition $|\lambda_\theta(x_m, \tau^n)| \le 1$ is equivalent to

$$|b_{1,m}^n e^{i\theta} + b_{0,m}^n + b_{-1,m}^n e^{-i\theta}|^2 - |a_{1,m}^n e^{i\theta} + a_{0,m}^n + a_{-1,m}^n e^{-i\theta}|^2 \le 0 \qquad (7.24)$$

if $|a_{1,m}^n e^{i\theta} + a_{0,m}^n + a_{-1,m}^n e^{-i\theta}|^2 \ge \tilde{c} > 0$, $\tilde{c}$ being a constant. The latter is easier to use in practice than the former.

Let us analyze the stability of scheme (7.6) in this way. This scheme has the form (7.23) with

$$\begin{aligned}
a_{1,m}^n &= -\left(\frac{a_m^{n+1/2}}{2\varDelta x^2} + \frac{b_m^{n+1/2}}{4\varDelta x}\right)\varDelta\tau,\\
a_{0,m}^n &= 1 + \frac{a_m^{n+1/2}}{\varDelta x^2}\varDelta\tau,\\
a_{-1,m}^n &= -\left(\frac{a_m^{n+1/2}}{2\varDelta x^2} - \frac{b_m^{n+1/2}}{4\varDelta x}\right)\varDelta\tau,\\
b_{1,m}^n &= -a_{1,m}^n,\\
b_{0,m}^n &= 2 - a_{0,m}^n,\\
b_{-1,m}^n &= -a_{-1,m}^n.
\end{aligned}$$

Here, we assume

$$g_m^{n+1/2} = c_m^{n+1/2} = 0$$

because we analyze the stability with respect to initial values only and ignoring a term of $O(\varDelta\tau)$ in coefficients will have no effect on the conclusion on stability. The left-hand side of the condition (7.24) for this scheme is

$$\begin{aligned}
&\left[-a_{1,m}^n e^{i\theta} + (2 - a_{0,m}^n) - a_{-1,m}^n e^{-i\theta}\right]\left[-a_{1,m}^n e^{-i\theta} + (2 - a_{0,m}^n) - a_{-1,m}^n e^{i\theta}\right]\\
&-(a_{1,m}^n e^{i\theta} + a_{0,m}^n + a_{-1,m}^n e^{-i\theta})(a_{1,m}^n e^{-i\theta} + a_{0,m}^n + a_{-1,m}^n e^{i\theta})\\
=\ &(a_{1,m}^n)^2 + (a_{0,m}^n - 2)^2 + (a_{-1,m}^n)^2 + 2a_{1,m}^n(a_{0,m}^n - 2)\cos\theta\\
&+2(a_{0,m}^n - 2)a_{-1,m}^n\cos\theta + 2a_{1,m}^n a_{-1,m}^n\cos 2\theta\\
&-\left[(a_{1,m}^n)^2 + (a_{0,m}^n)^2 + (a_{-1,m}^n)^2 + 2a_{1,m}^n a_{0,m}^n\cos\theta + 2a_{0,m}^n a_{-1,m}^n\cos\theta\right.\\
&\left.+2a_{1,m}^n a_{-1,m}^n\cos 2\theta\right]\\
=\ &(a_{0,m}^n - 2)^2 - (a_{0,m}^n)^2 - 4a_{1,m}^n\cos\theta - 4a_{-1,m}^n\cos\theta\\
=\ &-\frac{4a_m^{n+1/2}}{\varDelta x^2}\varDelta\tau + \frac{4a_m^{n+1/2}}{\varDelta x^2}\varDelta\tau\cos\theta\\
=\ &\frac{4a_m^{n+1/2}}{\varDelta x^2}\varDelta\tau(\cos\theta - 1).
\end{aligned}$$

This expression is always nonpositive. Therefore, the condition (7.24) is satisfied at every grid point. For scheme (7.6), there is no other boundary condition. Consequently, the scheme is expected to be stable.

So far, we say that a scheme is stable with respect to initial values if the error of the solution caused by the error in the initial condition is less than or equal to the error in the initial condition. However, generally speaking, we say that a scheme is stable with respect to initial values if the error of the solution caused by the error in the initial condition is less than c times the error in the initial condition. c is a constant independent of $\varDelta x$ and $\varDelta\tau$, but is allowed to be greater than one. That is, the error is allowed to increase by a certain factor, but the factor must be bounded and independent of $\varDelta x$ and $\varDelta\tau$. Therefore, we can take

$$|\lambda_\theta(x_m,\tau^n)| \le 1+\bar{c}\varDelta\tau \tag{7.25}$$

as a criterion for stability.[2] In fact, if the inequality (7.25) holds for any θ, then usually we can have

$$||\mathbf{e}^n||_{\mathrm{L}_2} \le (1+\bar{c}\varDelta\tau)||\mathbf{e}^{n-1}||_{\mathrm{L}_2} \le (1+\bar{c}\varDelta\tau)^n||\mathbf{e}^0||_{\mathrm{L}_2} \le \mathrm{e}^{\bar{c}nT/N}||\mathbf{e}^0||_{\mathrm{L}_2}$$

for any $n \le N$, so the error increases at most by a factor $\mathrm{e}^{\bar{c}T}$. Here we have used the relation $(1+\bar{c}\varDelta\tau)^{\frac{1}{\bar{c}\varDelta\tau}} \le \mathrm{e}$ for any positive $\varDelta\tau$.

Now let us study the stability of the difference scheme (7.5) by using the criterion (7.25). We consider the stability with respect to initial values only, so we can set $g_m^n = 0$. In this case, the scheme has the form (7.23) with $a_{1,m}^n = 0$, $a_{0,m}^n = 1$, $a_{-1,m}^n = 0$ and

$$\begin{aligned}
b_{1,m}^n &= \frac{a_m^n\varDelta\tau}{\varDelta x^2} + \frac{b_m^n\varDelta\tau}{2\varDelta x},\\
b_{0,m}^n &= 1 - 2\frac{a_m^n\varDelta\tau}{\varDelta x^2} + c_m^n\varDelta\tau,\\
b_{-1,m}^n &= \frac{a_m^n\varDelta\tau}{\varDelta x^2} - \frac{b_m^n\varDelta\tau}{2\varDelta x}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lambda_\theta(x_m,\tau^n) &= b_{1,m}^n\mathrm{e}^{\mathrm{i}\theta} + b_{0,m}^n + b_{-1,m}^n\mathrm{e}^{-\mathrm{i}\theta}\\
&= b_{0,m}^n + \left(b_{1,m}^n + b_{-1,m}^n\right)\cos\theta + \mathrm{i}\left(b_{1,m}^n - b_{-1,m}^n\right)\sin\theta\\
&= 1 - 2\frac{a_m^n\varDelta\tau}{\varDelta x^2} + c_m^n\varDelta\tau + 2\frac{a_m^n\varDelta\tau}{\varDelta x^2}\cos\theta + \mathrm{i}\frac{b_m^n\varDelta\tau}{\varDelta x}\sin\theta\\
&= 1 - 4\frac{a_m^n\varDelta\tau}{\varDelta x^2}\sin^2\frac{\theta}{2} + c_m^n\varDelta\tau + \mathrm{i}\frac{b_m^n\varDelta\tau}{\varDelta x}\sin\theta.
\end{aligned}$$

If

$$\max\frac{a_m^n\varDelta\tau}{\varDelta x^2} \le \frac{1}{2} \quad \text{or} \quad \frac{\varDelta\tau}{\varDelta x^2} \le \frac{1}{2\max a_m^n}, \tag{7.26}$$

[2] This criterion is equivalent to

$$|b_{1,m}^n\mathrm{e}^{\mathrm{i}\theta} + b_{0,m}^n + b_{-1,m}^n\mathrm{e}^{-\mathrm{i}\theta}|^2 - |a_{1,m}^n\mathrm{e}^{\mathrm{i}\theta} + a_{0,m}^n + a_{-1,m}^n\mathrm{e}^{-\mathrm{i}\theta}|^2 \le \bar{\bar{c}}\varDelta\tau$$

if $|a_{1,m}^n\mathrm{e}^{\mathrm{i}\theta} + a_{0,m}^n + a_{-1,m}^n\mathrm{e}^{-\mathrm{i}\theta}|^2 \ge \tilde{c} > 0$, $\tilde{c}$ being a constant, which is easier to use in practice than the criterion (7.25).

then

$$\begin{aligned}|\lambda_\theta(x_m,\tau^n)|^2 &\le (1+|c_m^n|\,\varDelta\tau)^2+\left(\frac{b_m^n\varDelta\tau}{\varDelta x}\right)^2\\ &\le (1+|c_m^n|\,\varDelta\tau)^2+\frac{(b_m^n)^2}{2\max a_m^n}\varDelta\tau\\ &\le (1+|c_m^n|\,\varDelta\tau)^2+2(1+|c_m^n|\,\varDelta\tau)\frac{(b_m^n)^2}{4\max a_m^n}\varDelta\tau\\ &\quad+\left[\frac{(b_m^n)^2}{4\max a_m^n}\varDelta\tau\right]^2\\ &=\left[1+|c_m^n|\,\varDelta\tau+\frac{(b_m^n)^2}{4\max a_m^n}\varDelta\tau\right]^2.\end{aligned}$$

Thus, let $\bar{c}=|c_m^n|+(b_m^n)^2/(4\max a_m^n)$, we have

$$|\lambda_\theta(x_m,\tau^n)|\le 1+\bar{c}\varDelta\tau$$

and we can expect this scheme to be stable if inequality (7.26) holds.

In fact, the stability of scheme (7.6) with variable coefficients has been proved rigorously in the paper [79] by Sun, Yan, and Zhu. By a similar method, the stability of scheme (7.5) with variable coefficients can also be shown when inequality (7.26) holds. If readers are interested in such a subject, please see that paper and the book [97] by Zhu, Zhong, Chen, and Zhang.

7.2.2 Convergence

If a scheme is stable with respect to initial values, and the truncation error of the scheme goes to zero as $\varDelta x$ and $\varDelta\tau$ tend to zero, then the approximate solution will usually go to the exact solution. Such a result is usually referred to as the Lax equivalence theorem (see the book [67] by Richtmyer and Morton). We are not going to prove this conclusion for general cases but explain this result intuitively through proving this result for special cases.

Consider the explicit finite-difference method (7.8). We know that the exact solution $u(x,\tau)$ satisfies the equation

$$\begin{aligned}&u(x_m,\tau^{n+1})\\ &=\alpha u(x_{m+1},\tau^n)+(1-2\alpha)u(x_m,\tau^n)+\alpha u(x_{m-1},\tau^n)+\varDelta\tau R_m^n(\varDelta x^2,\varDelta\tau),\\ &\qquad m=1,2,\cdots,M-1,\quad n=0,1,\cdots,N-1,\end{aligned}$$

where

$$R_m^n(\varDelta x^2,\varDelta\tau)=\frac{\varDelta\tau}{2}\frac{\partial^2 u}{\partial\tau^2}(x_m,\eta)-a\frac{\varDelta x^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi,\tau^n).$$

Let e_m^n be the error of the approximate solution on the point (x_m, τ^n), that is,

$$e_m^n = u(x_m, \tau^n) - u_m^n, \quad m = 0, 1, \cdots, M, \; n = 0, 1, \cdots, N.$$

Then, e_m^n is the solution of the problem

$$\begin{cases} e_m^{n+1} = \alpha e_{m+1}^n + (1-2\alpha)e_m^n + \alpha e_{m-1}^n + \varDelta\tau R_m^n(\varDelta x^2, \varDelta\tau), \\ \qquad\qquad\qquad m = 1, 2, \cdots, M-1, \qquad n = 0, 1, \cdots, N-1, \\ e_0^{n+1} = 0, \qquad\qquad\qquad\qquad\qquad\qquad n = 0, 1, \cdots, N-1, \\ e_M^{n+1} = 0, \qquad\qquad\qquad\qquad\qquad\qquad n = 0, 1, \cdots, N-1, \\ e_m^0 = 0, \qquad\qquad\qquad\qquad\qquad\qquad\; m = 0, 1, \cdots, M. \end{cases}$$

Because $e_0^n = e_M^n = 0$ for any n, the system can be written as

$$\begin{cases} \mathbf{e}^{n+1} = \mathbf{A}_1\mathbf{e}^n + \varDelta\tau\mathbf{R}^n(\varDelta x^2, \varDelta\tau), \quad n = 0, 1, \cdots, N-1, \\ \mathbf{e}^0 = 0, \end{cases}$$

where $\mathbf{e}^n$ is a vector with $M-1$ components e_m^n, $m = 1, 2, \cdots, M-1$ and

$$\mathbf{R}^n(\varDelta x^2, \varDelta\tau) = \begin{bmatrix} R_1^n(\varDelta x^2, \varDelta\tau) \\ R_2^n(\varDelta x^2, \varDelta\tau) \\ \vdots \\ R_{M-1}^n(\varDelta x^2, \varDelta\tau) \end{bmatrix}.$$

Actually, $\mathbf{e}^n$ can be written as $\sum_{k=1}^n \mathbf{e}_{(k)}^n$. Here, for $k = n$,

$$\mathbf{e}_{(n)}^n = \varDelta\tau\mathbf{R}^{n-1}(\varDelta x^2, \varDelta\tau)$$

and for $k = 1, 2, \cdots, n-1$, $\mathbf{e}_{(k)}^n$ is the solution of the following problem

$$\begin{cases} \mathbf{e}_{(k)}^{\bar{n}+1} = \mathbf{A}_1\mathbf{e}_{(k)}^{\bar{n}}, \quad \bar{n} = k, k+1, \cdots, n-1, \\ \mathbf{e}_{(k)}^k = \varDelta\tau\mathbf{R}^{k-1}(\varDelta x^2, \varDelta\tau). \end{cases}$$

Because the error does not increase for the scheme (7.8) if $\alpha \le 1/2$, $||\mathbf{e}^n||_{\mathrm{L}_2}$ should not be greater than $\sum_{k=1}^n \varDelta\tau||\mathbf{R}^{k-1}(\varDelta x^2, \varDelta\tau)||_{\mathrm{L}_2}$. Noticing $n \le T/\varDelta\tau$, we see that e_m^n goes to zero as $R_m^{k-1}(\varDelta x^2, \varDelta\tau)$ tends to zero for $k = 1, 2, \cdots, n$ and $m = 1, 2, \cdots, M-1$. Hence, the approximate solution converges to the exact solution as $\varDelta x$ and $\varDelta\tau$ tend to zero and α stays less than $1/2$ and $||\mathbf{e}^n||_{\mathrm{L}_2}$ has an order of $O(\varDelta x^2, \varDelta\tau)$. Usually, $\alpha = a\varDelta\tau/\varDelta x^2$ stays constant as $\varDelta x$ and $\varDelta\tau$ tend to zero. Therefore, $||\mathbf{e}^n||_{\mathrm{L}_2} = O(\varDelta\tau)$, and we say that the scheme (7.8) converges with order of $\varDelta\tau$.

For implicit schemes, the situation is similar. Consider the Crank–Nicolson scheme (7.9). The exact solution satisfies

$$\begin{aligned}&\frac{u(x_m,\tau^{n+1})-u(x_m,\tau^n)}{\Delta\tau}\\ &=\frac{a}{2}\left[\frac{u(x_{m+1},\tau^{n+1})-2u(x_m,\tau^{n+1})+u(x_{m-1},\tau^{n+1})}{\Delta x^2}\right.\\ &\quad\left.+\frac{u(x_{m+1},\tau^n)-2u(x_m,\tau^n)+u(x_{m-1},\tau^n)}{\Delta x^2}\right]+R_m^n(\Delta x^2,\Delta\tau^2),\\ &\qquad m=1,2,\cdots,M-1,\end{aligned}$$

where

$$\begin{aligned}&R_m^n(\Delta x^2,\Delta\tau^2)\\ &=\Delta\tau^2\left[\frac{1}{24}\frac{\partial^3 u}{\partial\tau^3}(x_m,\eta^{(1)})-\frac{a}{8}\frac{\partial^4 u}{\partial x^2\tau^2}(x_m,\eta^{(2)})\right]-\frac{\Delta x^2 a}{12}\frac{\partial^4 u}{\partial x^4}(\xi,\eta^{(3)}).\end{aligned}$$

In this case, the error satisfies

$$\mathbf{A}\mathbf{e}^{n+1}=\mathbf{B}\mathbf{e}^n+\Delta\tau\mathbf{R}^n(\Delta x^2,\Delta\tau^2),$$

where $\mathbf{e}^n$ and $\mathbf{R}^n(\Delta x^2,\Delta\tau^2)$ are two $(M-1)$-dimensional vectors with e_m^n and $R_m^n(\Delta x^2,\Delta\tau^2)$ as components, respectively, and $\mathbf{A}$ and $\mathbf{B}$ are given in the difference scheme (7.10). Just like in the case of the scheme (7.8), $\mathbf{e}^n$ can also be written as $\sum_{k=1}^n \mathbf{e}_{(k)}^n$. Here, for $k=n$,

$$\mathbf{e}_{(n)}^n=\Delta\tau\mathbf{A}^{-1}\mathbf{R}^{n-1}(\Delta x^2,\Delta\tau^2)$$

and for $k=1,2,\cdots,n-1$, $\mathbf{e}_{(k)}^n$ is the solution of the following problem:

$$\begin{cases}\mathbf{A}\mathbf{e}_{(k)}^{\bar{n}+1}=\mathbf{B}\mathbf{e}_{(k)}^{\bar{n}}, & \bar{n}=k,k+1,\cdots,n-1,\\ \mathbf{e}_{(k)}^k=\Delta\tau\mathbf{A}^{-1}\mathbf{R}^{k-1}(\Delta x^2,\Delta\tau^2).\end{cases}$$

The Crank–Nicolson scheme is stable with respect to the initial value. Thus, $||\mathbf{e}^n||_{\mathrm{L}_2}$ does not exceed $\sum_{k=1}^n \Delta\tau||\mathbf{A}^{-1}\mathbf{R}^{k-1}(\Delta x^2,\Delta\tau^2)||_{\mathrm{L}_2}$. Because

$$\mathbf{A}\mathbf{e}_{\omega_k}=\left(1+2\alpha\sin^2\frac{\omega_k}{2}\right)\mathbf{e}_{\omega_k},$$

we see that $1+2\alpha\sin^2(\omega_k/2)$ is an eigenvalue of $\mathbf{A}$. Thus, $1/[1+2\alpha\sin^2(\omega_k/2)]$ is an eigenvalue of $\mathbf{A}^{-1}$. This means that $\mathbf{A}^{-1}$ always exists and that its norm is bounded for any case. Consequently, $||\mathbf{e}^n||_{\mathrm{L}_2}$ goes to zero as Δx and $\Delta\tau$ tend to zero. In this case, we say that this scheme is convergent. Furthermore, because $||\mathbf{e}^n||_{\mathrm{L}_2}$ is of the order $O(\Delta x^2,\Delta\tau^2)$, we say that the scheme has a second-order convergence or possesses a second-order accuracy.

For schemes with variable coefficients, from the stability with respect to initial values and the consistency of a scheme, we also can have its convergence. Here, we say that a scheme is consistent with the partial differential equation if the truncation error of the scheme goes to zero as Δx and $\Delta\tau$ tend to zero. In the paper [79] by Sun, Yan, and Zhu, some results on this issue are given.

7.3 Extrapolation of Numerical Solutions

When a partial differential equation problem is discretized, a truncation error is introduced that causes the numerical solution to have an error. What is the relation between the truncation error and the error of the numerical solution? Intuitively, the answer should be that a term of $O(\Delta x^{k_1}, \Delta\tau^{k_2})$ in the truncation error causes an error of $O(\Delta x^{k_1}, \Delta\tau^{k_2})$ in the numerical solution. Here $O(\Delta x^{k_1}, \Delta\tau^{k_2})$ denotes a term less than $C\left(\Delta x^{k_1} + \Delta x^{k_2}\right)$, where C is a constant. Let us illustrate this fact.

Consider the following problem

$$\begin{cases} \dfrac{\partial u}{\partial \tau} = a(x,\tau)\dfrac{\partial^2 u}{\partial x^2} + b(x,\tau)\dfrac{\partial u}{\partial x} + c(x,\tau)u + g(x,\tau), \\ \qquad\qquad\qquad\quad 0 \le x \le 1, \quad 0 \le \tau \le T, \\ u(x,0) = f(x), \quad 0 \le x \le 1, \end{cases}$$

where $b(0,\tau) = a(0,\tau) = a_x(0,\tau) = b(1,\tau) = a(1,\tau) = a_x(1,\tau) = 0$ and $a(x,\tau) \ge 0$. This problem can be approximated by

$$\begin{cases} \delta_\tau u_m^{n+1/2} = a_m^{n+1/2}\delta_x^2 u_m^{n+1/2} + b_m^{n+1/2}\delta_{0x} u_m^{n+1/2} + c_m^{n+1/2} u_m^{n+1/2} + g_m^{n+1/2}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad 0 \le m \le M, \quad 0 \le n \le N-1, \\ u_m^0 = f(x_m), \qquad\qquad\qquad\qquad\quad\; 0 \le m \le M. \end{cases} \tag{7.27}$$

Here,

$$\begin{aligned} \delta_\tau u_m^{n+1/2} &= \frac{u_m^{n+1} - u_m^n}{\Delta\tau}, \\ \delta_x^2 u_m^{n+1/2} &= \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2} + \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2}\right), \\ \delta_{0x} u_m^{n+1/2} &= \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2\Delta x} + \frac{u_{m+1}^n - u_{m-1}^n}{2\Delta x}\right), \\ f_m^{n+1/2} &= \frac{1}{2}\left(f_m^{n+1} + f_m^n\right), \quad f \text{ being } u, a, b, c, g, \end{aligned}$$

and the same notation will be used for other functions in what follows. The truncation error of this scheme is $O(\Delta x^2) + O(\Delta\tau^2)$ everywhere; more accurately, it is in the form

$$P_m^{n+1/2}\Delta x^2 + R_m^{n+1/2}\Delta\tau^2 + O(\Delta x^4 + \Delta\tau^4),$$

where $P_m^{n+1/2}$ and $R_m^{n+1/2}$ denote the values of two functions $P(x,\tau)$ and $R(x,\tau)$ at $x = x_m$ and $\tau = \tau^{n+1/2}$. That is, the exact solution satisfies the following equation:

$$\begin{cases} \delta_\tau U_m^{n+1/2} = a_m^{n+1/2}\delta_x^2 U_m^{n+1/2} + b_m^{n+1/2}\delta_{0x}U_m^{n+1/2} + c_m^{n+1/2}U_m^{n+1/2} + g_m^{n+1/2} \\ \qquad + P_m^{n+1/2}\Delta x^2 + R_m^{n+1/2}\Delta\tau^2 + O(\Delta x^4 + \Delta\tau^4), \\ \qquad\qquad\qquad\qquad 0 \le m \le M, \quad 0 \le n \le N-1, \\ u_m^0 = f(x_m), \qquad\qquad 0 \le m \le M, \end{cases}$$

where U_m^n stands for $u(x_m, \tau^n)$. Suppose v_1 and v_2 are the solutions of the problems

$$\begin{cases} \dfrac{\partial v_1}{\partial \tau} = a(x,\tau)\dfrac{\partial^2 v_1}{\partial x^2} + b(x,\tau)\dfrac{\partial v_1}{\partial x} + c(x,\tau)v_1 + P(x,\tau), \\ \qquad\qquad\qquad 0 \le x \le 1, \quad 0 \le \tau \le T, \\ v_1(x,0) = 0, \qquad\qquad 0 \le x \le 1 \end{cases}$$

and

$$\begin{cases} \dfrac{\partial v_2}{\partial \tau} = a(x,\tau)\dfrac{\partial^2 v_2}{\partial x^2} + b(x,\tau)\dfrac{\partial v_2}{\partial x} + c(x,\tau)v_2 + R(x,\tau), \\ \qquad\qquad\qquad 0 \le x \le 1, \quad 0 \le \tau \le T, \\ v_2(x,0) = 0, \qquad\qquad 0 \le x \le 1, \end{cases}$$

respectively. Let $V_{1,m}^n$ and $V_{2,m}^n$ denote $v_1(x_m, \tau^n)$ and $v_2(x_m, \tau^n)$. Then,

$$\begin{cases} \delta_\tau V_{1,m}^{n+1/2} = a_m^{n+1/2}\delta_x^2 V_{1,m}^{n+1/2} + b_m^{n+1/2}\delta_{0x}V_{1,m}^{n+1/2} + c_m^{n+1/2}V_{1,m}^{n+1/2} + P_m^{n+1/2} \\ \qquad + O(\Delta x^2 + \Delta\tau^2), \qquad 0 \le m \le M, \quad 0 \le n \le N-1, \\ V_{1,m}^0 = 0, \qquad\qquad 0 \le m \le M, \end{cases}$$

and

$$\begin{cases} \delta_\tau V_{2,m}^{n+1/2} = a_m^{n+1/2}\delta_x^2 V_{2,m}^{n+1/2} + b_m^{n+1/2}\delta_{0x}V_{2,m}^{n+1/2} + c_m^{n+1/2}V_{2,m}^{n+1/2} + R_m^{n+1/2} \\ \qquad + O(\Delta x^2 + \Delta\tau^2), \qquad 0 \le m \le M, \quad 0 \le n \le N-1, \\ V_{2,m}^0 = 0, \qquad\qquad 0 \le m \le M. \end{cases}$$

Let us define

$$W_m^n = U_m^n - u_m^n - V_{1,m}^n \Delta x^2 - V_{2,m}^n \Delta\tau^2.$$

It is clear that W_m^n satisfies

$$\begin{cases} \delta_\tau W_m^{n+1/2} = a_m^{n+1/2}\delta_x^2 W_m^{n+1/2} + b_m^{n+1/2}\delta_{0x}W_m^{n+1/2} + c_m^{n+1/2}W_m^{n+1/2} \\ \qquad + O(\Delta x^4 + \Delta x^2\Delta\tau^2 + \Delta\tau^4), \quad 0 \le m \le M, \quad 0 \le n \le N-1, \\ W_m^0 = 0, \qquad\qquad 0 \le m \le M. \end{cases}$$

Because the scheme is stable with respect to the initial value and the nonhomogeneous term (see the paper [76] by Sun and the paper [79] by Sun, Yan, and Zhu for the details of the proof) and $O(\Delta x^2 \Delta\tau^2)$ can be expressed as $O(\Delta x^4 + \Delta\tau^4)$, we have

$$|U_m^n - u_m^n - V_{1,m}^n \Delta x^2 - V_{2,m}^n \Delta\tau^2| \le O(\Delta x^4 + \Delta\tau^4),$$

or we can write this relation as

$$u(x_m,\tau^n)-u_m^n(\Delta x,\Delta\tau)=v_1(x_m,\tau^n)\Delta x^2+v_2(x_m,\tau^n)\Delta\tau^2+O(\Delta x^4+\Delta\tau^4),$$

that is,

$$\begin{aligned}u_m^n(\Delta x,\Delta\tau)=u(x_m,\tau^n)&-v_1(x_m,\tau^n)\Delta x^2-v_2(x_m,\tau^n)\Delta\tau^2\\&+O(\Delta x^4+\Delta\tau^4).\end{aligned}\tag{7.28}$$

Here, we write u_m^n as $u_m^n(\Delta x,\Delta\tau)$ in order to indicate that the approximate solution is obtained on a mesh with mesh sizes Δx and $\Delta\tau$. For this case, the error of a numerical solution is in the form

$$v_1(x_m,\tau^n)\Delta x^2+v_2(x_m,\tau^n)\Delta\tau^2+O(\Delta x^4+\Delta\tau^4),$$

which has the same form as the truncation error given above. Similarly, if the truncation error of a numerical scheme, including the algorithms for boundary conditions, is

$$P\Delta x^2+Q\Delta x\Delta\tau+R\Delta\tau^2+O(\Delta\tau^3),$$

i.e., the scheme is second order and stable, then the numerical solution can be expressed as

$$\begin{aligned}u_m^n(\Delta x,\Delta\tau)=u(x_m,\tau^n)&-v_1(x_m,\tau^n)\Delta x^2-v_{12}(x_m,\tau^n)\Delta x\Delta\tau\\&-v_2(x_m,\tau^n)\Delta x^2+O(\Delta\tau^3),\end{aligned}\tag{7.29}$$

where $O(\Delta\tau^3)$ means $O(\Delta x^3+\Delta x^2\Delta\tau+\Delta x\Delta\tau^2+\Delta\tau^3)$ for simplicity.

Here, the approximate value is given only at the nodes. Now let us generate a function defined on the domain $[0,1]\times[0,T]$ by some type of interpolation. We assume that the interpolation function generated from the values on the nodes by an interpolation method is an approximation to $f(x,\tau)$ with an error of $O(\Delta\tau^3)$ for any smooth enough function $f(x,\tau)$. For example, if we use quadratic interpolation, then the interpolation function generated has such a property. Let $u(x,\tau;\Delta x,\Delta\tau)$ denote such a function generated by $u(x_m,\tau^n;\Delta x,\Delta\tau)$. Because $u(x_m,\tau^n;\Delta x,\Delta\tau)$ consists of $u(x_m,\tau^n)-v_1(x_m,\tau^n)\Delta x^2-v_{12}(x_m,\tau^n)\Delta x\Delta\tau-v_2(x_m,\tau^n)\Delta\tau^2$ and $O(\Delta\tau^3)$, the interpolation function also has two parts. One part is the interpolation function generated by $u(x_m,\tau^n)-v_1(x_m,\tau^n)\Delta x^2-v_{12}(x_m,\tau^n)\Delta x\Delta\tau-v_2(x_m,\tau^n)\Delta\tau^2$, which we call $u_1(x,\tau;\Delta x,\Delta\tau)$. The other part is generated by the term $O(\Delta\tau^3)$, which is denoted by $u_2(x,\tau;\Delta x,\Delta\tau)$. Clearly,

$$u_1(x,\tau;\Delta x,\Delta\tau)-u(x,\tau)+v_1(x,\tau)\Delta x^2+v_{12}(x,\tau)\Delta x\Delta\tau+v_2(x,\tau)\Delta\tau^2$$

is a term of $O(\Delta\tau^3)$. The function $u_2(x,\tau;\Delta x,\Delta\tau)$ is also a term of $O(\Delta\tau^3)$. Consequently, we have

$$\begin{aligned}u(x,\tau;\Delta x,\Delta\tau) &= u_1(x,\tau;\Delta x,\Delta\tau)+u_2(x,\tau;\Delta x,\Delta\tau)\\ &= u(x,\tau)-v_1(x,\tau)\Delta x^2-v_{12}(x,\tau)\Delta x\Delta\tau-v_2(x,\tau)\Delta\tau^2\\ &\quad +O(\Delta\tau^3).\end{aligned}$$

In this case, we can use the following technique to eliminate the error of $O(\Delta x^2+\Delta x\Delta\tau+\Delta\tau^2)$ if we have numerical solutions on a mesh with mesh sizes Δx and $\Delta\tau$ and on a mesh with mesh sizes $2\Delta x$ and $2\Delta\tau$. Let us consider a linear combination of the solutions on the two different meshes, which are denoted by $u(x,\tau;\Delta x,\Delta\tau)$ and $u(x,\tau;\,2\Delta x,2\Delta\tau)$:

$$\begin{aligned}&(1-d)\times u(x,\tau;\Delta x,\Delta\tau)+d\times u(x,\tau;2\Delta x,2\Delta\tau)\\ =\;& u(x,\tau)-v_1(x,\tau)(1-d+4d)\Delta x^2-v_{12}(x,\tau)(1-d+4d)\Delta x\Delta\tau\\ &-v_2(x,\tau)(1-d+4d)\Delta\tau^2+O(\Delta\tau^3).\end{aligned}$$

If we choose d such that $1-d+4d=0$, that is, $d=-\dfrac{1}{3}$, then

$$(1-d)\times u(x,\tau;\Delta x,\Delta\tau)+d\times u(x,\tau;2\Delta x,2\Delta\tau)=u(x,\tau)+O(\Delta\tau^3).$$

Therefore,

$$\frac{1}{3}[4u(x,\tau;\Delta x,\Delta\tau)-u(x,\tau;2\Delta x,2\Delta\tau)] \tag{7.30}$$

is an approximate to $u(x,\tau)$ with an error of $O(\Delta\tau^3)$.

However, for the approximation (7.27), the expression of the numerical solution is in the form (7.28), and the extrapolation formula of numerical solutions (7.30) gives an approximation to $u(x,\tau)$ with an error of $O(\Delta\tau^4)$. This is a special case. Generally speaking, if for a second-order scheme we have three solutions $u_m^n(\Delta x,\Delta\tau), u_m^n(2\Delta x,2\Delta\tau)$, and $u_m^n(4\Delta x,4\Delta\tau)$, then we can have an approximation with an error of $O(\Delta\tau^4)$. In order to do that, we first generate an interpolation function from the values at these nodes and require the interpolation with an error of $O(\Delta\tau^4)$. This can be done, for example, by cubic interpolation. Let $u(x,\tau;\Delta x,\Delta\tau)$, $u(x,\tau;2\Delta x,2\Delta\tau)$, and $u(x,\tau;4\Delta x,4\Delta\tau)$ represent these functions. Then, consider a linear combination of them:

$$(1-d_1-d_2)u(x,\tau;\Delta x,\Delta\tau)+d_1u(x,\tau;2\Delta x,2\Delta\tau)+d_2u(x,\tau;4\Delta x,4\Delta\tau).$$

If we choose $d_1$and d_2 such that

$$\begin{cases}1-d_1-d_2+2^2d_1+4^2d_2=0,\\ 1-d_1-d_2+2^3d_1+4^3d_2=0,\end{cases}$$

which gives

$$\begin{cases}d_1=-\dfrac{12}{21},\\[2ex] d_2=\dfrac{1}{21},\end{cases}$$

then all the terms of $O(\varDelta\tau^2)$ and the terms of $O(\varDelta\tau^3)$ in

$$(1-d_1-d_2)u(x,\tau;\varDelta x,\varDelta\tau)+d_1u(x,\tau;2\varDelta x,2\varDelta\tau)+d_2u(x,\tau;4\varDelta x,4\varDelta\tau)$$

are eliminated. Therefore

$$\frac{1}{21}[32u(x,\tau;\varDelta x,\varDelta\tau)-12u(x,\tau;2\varDelta x,2\varDelta\tau)+u(x,\tau;4\varDelta x,4\varDelta\tau)] \qquad (7.31)$$

gives an approximation to $u(x,\tau)$ with an error of $O(\varDelta\tau^4)$ for any second-order scheme.

Here, we need to point out that in order to obtain an approximate solution with an error of $O(\varDelta\tau^3)$, it is not necessary for both $\varDelta x_1/\varDelta x_2$ and $\varDelta\tau_1/\varDelta\tau_2$ to equal two, where $\varDelta x_1, \varDelta\tau_1$ are mesh sizes for one mesh and $\varDelta x_2, \varDelta\tau_2$ for the other. For example, if we have a solution on a 12×16 mesh and a solution on a 9×12 mesh, then we still can obtain an approximate solution with an error of $O(\varDelta\tau^3)$ by using extrapolation. Furthermore, if there exist solutions on 15×20, 12×16, and 9×12 meshes, then we can have an approximate solution with an error of $O(\varDelta\tau^4)$ by using extrapolation. These are left as a problem for the reader to prove. Generally speaking, when a scheme has an error of $\varDelta x^{k_1}$ and $\varDelta\tau^{k_2}$ and we know solutions on two meshes, the extrapolation can be used if $\dfrac{\varDelta x_1^{k_1}}{\varDelta\tau_1^{k_2}}=\dfrac{\varDelta x_2^{k_1}}{\varDelta\tau_2^{k_2}}$, where $\varDelta x_i$ and $\varDelta\tau_i$, $i=1,2$, are mesh sizes used in order to obtain the two solutions. For example, if $k_1=2$ and $k_2=1$, then when solutions on a 20×20 mesh and a 40×80 mesh are obtained, this technique can also be used because $\dfrac{\left(\frac{1}{20}\right)^2}{\frac{1}{20}}=\dfrac{\left(\frac{1}{40}\right)^2}{\frac{1}{80}}$ (see Problem 16).

The technique of generating more accurate results by combining several numerical results, which is similar to Richardson's extrapolation in numerical methods for ordinary differential equations, is referred to as the extrapolation technique of numerical solutions in next few chapters. Finally we need to point out that this technique works if the solution is smooth, but may not work if the solution is not smooth enough.

7.4 Two-Dimensional Degenerate Parabolic Equations

Generally speaking, the coefficients of PDEs are variable, and so the difference equations also have variable coefficients. For such a case, the theoretical analysis of numerical methods is more complicated. In this section, for some type of two-dimensional degenerate parabolic equations and for a special but popular scheme, a complete theoretical analysis of numerical methods is given.

Consider the following two-dimensional degenerate parabolic partial differential equation:

$$\frac{\partial u}{\partial \tau} = a_{11}(x,y,\tau)\frac{\partial^2 u}{\partial x^2} + 2a_{12}(x,y,\tau)\frac{\partial^2 u}{\partial x \partial y} + a_{22}(x,y,\tau)\frac{\partial^2 u}{\partial y^2} + b_1(x,y,\tau)\frac{\partial u}{\partial x}$$
$$+b_2(x,y,\tau)\frac{\partial u}{\partial y}+c(x,y,\tau)u+g(x,y,\tau), \quad (x,y)\in\Omega,\ 0\leq\tau\leq T, \tag{7.32}$$

with the initial condition

$$u(x,y,0) = f(x,y), \quad (x,y) \in \Omega, \tag{7.33}$$

where

$$\Omega = \{(x,y) \mid x_l \leq x \leq x_u, y_l \leq y \leq y_u\},$$

$$a_{11}(x,y,\tau)\Big|_{x=x_l \text{ or } x_u} = 0, \qquad a_{22}(x,y,\tau)\Big|_{y=y_l \text{ or } y_u} = 0, \tag{7.34}$$

$$b_1(x,y,\tau)\Big|_{x=x_l \text{ or } x_u} = 0, \qquad b_2(x,y,\tau)\Big|_{y=y_l \text{ or } y_u} = 0, \tag{7.35}$$

$$\frac{\partial a_{11}(x,y,\tau)}{\partial x}\Big|_{x=x_l \text{ or } x_u} = 0, \qquad \frac{\partial a_{22}(x,y,\tau)}{\partial y}\Big|_{y=y_l \text{ or } ,y_u} = 0, \tag{7.36}$$

and the matrix

$$\begin{pmatrix} a_{11}(x,y,\tau) & a_{12}(x,y,\tau) \\ a_{12}(x,y,\tau) & a_{22}(x,y,\tau) \end{pmatrix}$$

is semi-positive (nonnegative); i.e., for any $X \in \mathcal{R}$ and $Y \in \mathcal{R}$, we have

$$a_{11}(x,y,\tau)X^2 + 2a_{12}(x,y,\tau)XY + a_{22}(x,y,\tau)Y^2 \geq 0. \tag{7.37}$$

The matrix of the coefficients of second derivatives is semi-positive, so $a_{12}^2 \leq a_{11}a_{22}$. Thus, when $a_{11} = 0$ or $a_{22} = 0$, we have $a_{12} = 0$. Thus, from the expression (7.34), we have

$$a_{12}(x,y,\tau)\Big|_{x=x_l \text{ or } x_u} = 0, \qquad a_{12}(x,y,\tau)\Big|_{y=y_l \text{ or } y_u} = 0. \tag{7.38}$$

Taking the partial derivative of the first and second relations in the result (7.38) with respect to y and x, respectively, we can further have

$$\frac{\partial a_{12}(x,y,\tau)}{\partial y}\Big|_{x=x_l \text{ or } x_u} = 0, \qquad \frac{\partial a_{12}(x,y,\tau)}{\partial x}\Big|_{y=y_l \text{ or } y_u} = 0. \tag{7.39}$$

Denote

$$c_1 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|\frac{\partial^2 a_{11}(x,y,\tau)}{\partial x^2}\right|, \qquad c_2 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|\frac{\partial^2 a_{12}(x,y,\tau)}{\partial x \partial y}\right|,$$

$$c_3 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|\frac{\partial^2 a_{22}(x,y,\tau)}{\partial y^2}\right|, \qquad c_4 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|\frac{\partial b_1(x,y,\tau)}{\partial x}\right|,$$

$$c_5 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|\frac{\partial b_2(x,y,\tau)}{\partial y}\right|, \qquad c_6 = \max_{(x,y,\tau)\in\Omega\times[0,T]}\left|c(x,y,\tau)\right|,$$

and set

$$c = c_1 + 2c_2 + c_3 + c_4 + c_5 + 2c_6. \tag{7.40}$$

In Sect. 2.4.3, for more general problems we have obtained the following inequality:

$$\begin{aligned}\iint_{\Omega} u^2(x, y, \tau)dxdy \le e^{\bar{c}T} \Bigg[\iint_{\Omega} f^{\,2}(x, y)dxdy \\ + \int_0^{\tau} \Big(\iint_{\Omega} g^2(x, y, s)dxdy\Big)ds\Bigg], \quad 0 \le \tau \le T,\end{aligned}$$

where $\bar{c}$ is a constant determined by the bounds of the coefficients of the PDE and their derivatives. Of course, for the problem here, such an inequality holds. In this section, we are going to prove that for the numerical solutions obtained by a special but popular scheme, such an inequality still holds.

7.4.1 The Crank–Nicolson Difference Scheme and a Preliminary Lemma

Take three positive integers M, N, and K. Set $h_1 = (x_u - x_l)/M, h_2 = (y_u - y_l)/N, \Delta\tau = T/K$, and denote

$$\begin{aligned}&x_m = x_l + mh_1, \quad 0 \le m \le M,\\ &y_n = y_l + nh_2, \quad 0 \le n \le N,\\ &\tau^k = k\Delta\tau, \quad 0 \le k \le K,\\ &\Omega_h = \{(x_m, y_n) \mid 0 \le m \le M, 0 \le n \le N\},\\ &\Omega_{\Delta\tau} = \{\tau^k \mid 0 \le k \le K\}.\end{aligned}$$

Let $\mathcal{V} = \{u \mid u = \{u_{mn}, 0 \le m \le M, 0 \le n \le N\}\}$ be the grid function space on Ω_h. If $u \in \mathcal{V}$, we introduce the following notation:

$$\begin{aligned}&\delta_x u_{m+\frac{1}{2},n} = \tfrac{1}{h_1}(u_{m+1,n} - u_{mn}), && \Delta_x u_{mn} = \tfrac{1}{2h_1}(u_{m+1,n} - u_{m-1,n}),\\ &\delta_y u_{m,n+\frac{1}{2}} = \tfrac{1}{h_2}(u_{m,n+1} - u_{mn}), && \Delta_y u_{mn} = \tfrac{1}{2h_2}(u_{m,n+1} - u_{m,n-1}),\\ &\delta_x^2 u_{mn} = \tfrac{1}{h_1^2}(u_{m+1,n} - 2u_{mn} + u_{m-1,n}),\\ &\delta_y^2 u_{mn} = \tfrac{1}{h_2^2}(u_{m,n+1} - 2u_{mn} + u_{m,n-1}).\end{aligned}$$

It is obvious that

$$\begin{aligned}\Delta_x u_{mn} &= \frac{1}{2}(\delta_x u_{m+\frac{1}{2},n} + \delta_x u_{m-\frac{1}{2},n}),\ \delta_x^2 u_{mn} = \frac{1}{h_1}(\delta_x u_{m+\frac{1}{2},n} - \delta_x u_{m-\frac{1}{2},n}),\\ \Delta_y u_{mn} &= \frac{1}{2}(\delta_y u_{m,n+\frac{1}{2}} + \delta_y u_{m,n-\frac{1}{2}}),\ \delta_y^2 u_{mn} = \frac{1}{h_2}(\delta_y u_{m,n+\frac{1}{2}} - \delta_y u_{m,n-\frac{1}{2}}).\end{aligned}$$

For any $u \in \mathcal{V}$, and $v \in \mathcal{V}$, their inner product is defined by

$$(u,v) = h_1h_2\left[\sum_{m=1}^{M-1}\sum_{n=1}^{N-1} u_{mn}v_{mn} + \frac{1}{2}\sum_{m=1}^{M-1}(u_{m0}v_{m0} + u_{mN}v_{mN})\right.$$
$$\left. +\frac{1}{2}\sum_{n=1}^{N-1}(u_{0n}v_{0n} + u_{Mn}v_{Mn}) + \frac{1}{4}(u_{00}v_{00} + u_{M0}v_{M0} + u_{0N}v_{0N} + u_{MN}v_{MN})\right] \tag{7.41}$$

and the norm of a grid function is defined by

$$\|u\| = \sqrt{(u,u)}.$$

It is also obvious that the definition of the inner product can also be written in another form:

$$(u,v) = \frac{1}{4}h_1h_2\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\Big(u_{mn}v_{mn} + u_{m+1,n}v_{m+1,n}$$
$$+u_{m,n+1}v_{m,n+1} + u_{m+1,n+1}v_{m+1,n+1}\Big). \tag{7.42}$$

We also define the grid function U on $\Omega_h \times \Omega_{\Delta\tau}$ as follows:

$$U_{mn}^k = u(x_m, y_n, \tau^k), \quad 0 \le m \le M, \quad 0 \le n \le N, \quad 0 \le k \le K.$$

In what follows, we use the following notations:

$$U_{mn}^{k+\frac{1}{2}} = \frac{1}{2}(U_{mn}^{k+1} + U_{mn}^k), \quad \tau^{k+\frac{1}{2}} = \frac{1}{2}(\tau^k + \tau^{k+1})$$

and

$$\begin{aligned}
&(a_{11})_{mn}^{k+\frac{1}{2}} = a_{11}(x_m, y_n, \tau^{k+\frac{1}{2}}), && (a_{12})_{mn}^{k+\frac{1}{2}} = a_{12}(x_m, y_n, \tau^{k+\frac{1}{2}}),\\
&(a_{22})_{mn}^{k+\frac{1}{2}} = a_{22}(x_m, y_n, \tau^{k+\frac{1}{2}}), && (b_1)_{mn}^{k+\frac{1}{2}} = b_1(x_m, y_n, \tau^{k+\frac{1}{2}}),\\
&(b_2)_{mn}^{k+\frac{1}{2}} = b_2(x_m, y_n, \tau^{k+\frac{1}{2}}), && c_{mn}^{k+\frac{1}{2}} = c(x_m, y_n, \tau^{k+\frac{1}{2}}),\\
&g_{mn}^{k+\frac{1}{2}} = g(x_m, y_n, \tau^{k+\frac{1}{2}}), && f_{mn} = f(x_m, y_n).
\end{aligned}$$

Suppose problem (7.32)–(7.33) has a smooth solution $u(x,y,\tau)$. Applying the Taylor expansion, we can obtain

$$\frac{1}{\Delta\tau}(U_{mn}^{k+1} - U_{mn}^k) = (a_{11})_{mn}^{k+\frac{1}{2}}\delta_x^2 U_{mn}^{k+\frac{1}{2}} + 2(a_{12})_{mn}^{k+\frac{1}{2}}\Delta_x\Delta_y U_{mn}^{k+\frac{1}{2}}$$
$$+(a_{22})_{mn}^{k+\frac{1}{2}}\delta_y^2 U_{mn}^{k+\frac{1}{2}} + (b_1)_{mn}^{k+\frac{1}{2}}\Delta_x U_{mn}^{k+\frac{1}{2}} + (b_2)_{mn}^{k+\frac{1}{2}}\Delta_y U_{mn}^{k+\frac{1}{2}}$$
$$+c_{mn}^{k+\frac{1}{2}}U_{mn}^{k+\frac{1}{2}} + g_{mn}^{k+\frac{1}{2}} + R_{mn}^{k+\frac{1}{2}},$$
$$0 \le m \le M, \quad 0 \le n \le N, \quad 0 \le k \le K-1 \tag{7.43}$$

and there exists a constant c_0 such that

$$\begin{aligned}&|R_{mn}^{k+\frac{1}{2}}| \le c_0(h_1^2+h_2^2+\varDelta\tau^2),\\ &0\le m\le M,\quad 0\le n\le N,\quad 0\le k\le K-1.\end{aligned}\tag{7.44}$$

Omitting the small term $R_{mn}^{k+\frac{1}{2}}$ in the expression (7.43) and writing down the initial condition on $\varOmega_h$:

$$U_{mn}^0 = f_{mn},\quad 0\le m\le M,\quad 0\le n\le N,\tag{7.45}$$

we have for the problem (7.32)–(7.33) the following difference scheme:

$$\begin{aligned}\frac{1}{\varDelta\tau}(u_{mn}^{k+1}-u_{mn}^{k}) &= (a_{11})_{mn}^{k+\frac{1}{2}}\delta_x^2 u_{mn}^{k+\frac{1}{2}} + 2(a_{12})_{mn}^{k+\frac{1}{2}}\varDelta_x\varDelta_y u_{mn}^{k+\frac{1}{2}}\\ &\quad +(a_{22})_{mn}^{k+\frac{1}{2}}\delta_y^2 u_{mn}^{k+\frac{1}{2}} + (b_1)_{mn}^{k+\frac{1}{2}}\varDelta_x u_{mn}^{k+\frac{1}{2}} + (b_2)_{mn}^{k+\frac{1}{2}}\varDelta_y u_{mn}^{k+\frac{1}{2}} + c_{mn}^{k+\frac{1}{2}}u_{mn}^{k+\frac{1}{2}}\\ &\quad +g_{mn}^{k+\frac{1}{2}},\quad 0\le m\le M,\quad 0\le n\le N,\quad 0\le k\le K-1,\end{aligned}\tag{7.46}$$

$$u_{mn}^0 = f_{mn},\quad 0\le m\le M,\quad 0\le n\le N.\tag{7.47}$$

The following lemma will be used for the analysis of the difference scheme.

Lemma 7.1. *Let $u\in\mathcal{V}$. Then we have*

$$\begin{aligned}&\left(a_{11}^{k+\frac{1}{2}}\delta_x^2 u, u\right) + 2\left(a_{12}^{k+\frac{1}{2}}\varDelta_x\varDelta_y u, u\right) + \left(a_{22}^{k+\frac{1}{2}}\delta_y^2 u, u\right)\\ &+\left(b_1^{k+\frac{1}{2}}\varDelta_x u, u\right) + \left(b_2^{k+\frac{1}{2}}\varDelta_y u, u\right) + \left(c^{k+\frac{1}{2}}u, u\right) \le \frac{c}{2}\|u\|^2,\end{aligned}\tag{7.48}$$

where c is defined by the expression (7.40).

Section 7.4.2 is devoted to the proof of this lemma.

7.4.2 ‡The Proof of the Preliminary Lemma

We will estimate each term in the inequality (7.48). For simplicity, we omit the superscript.

Proposition 7.1 *For $\left(a_{11}\delta_x^2 u, u\right)$ and $\left(a_{22}\delta_y^2 u, u\right)$, we have the following inequalities:*

$$
\begin{aligned}
B_1 &\equiv \left(a_{11}\delta_x^2 u, u\right) \\
&\le -h_1 h_2 \left[\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{11})_{mn}\frac{(\delta_x u_{m-\frac{1}{2},n})^2+(\delta_x u_{m+\frac{1}{2},n})^2}{2}\right. \\
&\qquad +\frac{1}{2}\sum_{m=1}^{M-1}(a_{11})_{m0}\frac{(\delta_x u_{m-\frac{1}{2},0})^2+(\delta_x u_{m+\frac{1}{2},0})^2}{2} \\
&\qquad \left.+\frac{1}{2}\sum_{m=1}^{M-1}(a_{11})_{mN}\frac{(\delta_x u_{m-\frac{1}{2},N})^2+(\delta_x u_{m+\frac{1}{2},N})^2}{2}\right]+\frac{1}{2}c_1\|u\|^2 \\
&\le -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}\left[(a_{11})_{mn}\frac{(\delta_x u_{m-\frac{1}{2},n})^2+(\delta_x u_{m+\frac{1}{2},n})^2}{2}\right]+\frac{1}{2}c_1\|u\|^2.
\end{aligned}
\tag{7.49}
$$

and

$$
\begin{aligned}
B_3 &\equiv \left(a_{22}\delta_y^2 u, u\right) \\
&\le -h_1 h_2 \left[\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{22})_{mn}\frac{(\delta_y u_{m,n-\frac{1}{2}})^2+(\delta_y u_{m,n+\frac{1}{2}})^2}{2}\right. \\
&\qquad +\frac{1}{2}\sum_{n=1}^{N-1}(a_{22})_{0n}\frac{(\delta_y u_{0,n-\frac{1}{2}})^2+(\delta_y u_{0,n+\frac{1}{2}})^2}{2} \\
&\qquad \left.+\frac{1}{2}\sum_{n=1}^{N-1}(a_{22})_{Mn}\frac{(\delta_y u_{M,n-\frac{1}{2}})^2+(\delta_y u_{M,n+\frac{1}{2}})^2}{2}\right]+\frac{1}{2}c_3\|u\|^2 \\
&\le -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}\left[(a_{22})_{mn}\frac{(\delta_y u_{m,n-\frac{1}{2}})^2+(\delta_y u_{m,n+\frac{1}{2}})^2}{2}\right]+\frac{1}{2}c_3\|u\|^2.
\end{aligned}
\tag{7.50}
$$

Proof. Because $(a_{11})_{0n} = (a_{11})_{Mn} = 0$ for $n = 0, 1, \cdots, N$, some terms in the inner product are zero. Thus, the expression of $\left(a_{11}\delta_x^2 u, u\right)$ is

$$
\begin{aligned}
B_1 = \left(a_{11}\delta_x^2 u, u\right) = h_1 h_2 &\left[\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{11})_{mn}\ \delta_x^2 u_{mn}\ u_{mn}\right. \\
&\left.+\frac{1}{2}\sum_{m=1}^{M-1}(a_{11})_{m0}\ \delta_x^2 u_{m0}\ u_{m0}+\frac{1}{2}\sum_{m=1}^{M-1}(a_{11})_{mN}\ \delta_x^2 u_{mn}\ u_{mN}\right].
\end{aligned}
\tag{7.51}
$$

Averaging the following two equalities:

$$\begin{aligned}
&h_1 \sum_{m=1}^{M-1} (a_{11})_{mn}\, \delta_x^2 u_{mn}\, u_{mn} \\
&= \sum_{m=1}^{M-1} (a_{11})_{mn} (\delta_x u_{m+\frac{1}{2},n} - \delta_x u_{m-\frac{1}{2},n}) u_{mn} \\
&= \sum_{m=1}^{M-1} (a_{11})_{mn}\, \delta_x u_{m+\frac{1}{2},n}\, u_{mn} - \sum_{m=0}^{M-2} (a_{11})_{m+1,n}\, \delta_x u_{m+\frac{1}{2},n}\, u_{m+1,n} \\
&= \sum_{m=0}^{M-1} (a_{11})_{mn}\, \delta_x u_{m+\frac{1}{2},n}\, u_{mn} - \sum_{m=0}^{M-1} (a_{11})_{m+1,n}\, \delta_x u_{m+\frac{1}{2},n}\, u_{m+1,n} \\
&= \sum_{m=0}^{M-1} (a_{11})_{mn}\, \delta_x u_{m+\frac{1}{2},n}\, (u_{mn} - u_{m+1,n}) \\
&\quad + \sum_{m=0}^{M-1} [(a_{11})_{mn} - (a_{11})_{m+1,n}]\, \delta_x u_{m+\frac{1}{2},n}\, u_{m+1,n} \\
&= -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} (\delta_x u_{m+\frac{1}{2},n})^2 - h_1 \sum_{m=0}^{M-1} (\delta_x a_{11})_{m+\frac{1}{2},n}\, \delta_x u_{m+\frac{1}{2},n}\, u_{m+1,n}
\end{aligned}$$

and

$$\begin{aligned}
&h_1 \sum_{m=1}^{M-1} (a_{11})_{mn}\, \delta_x^2 u_{mn}\, u_{mn} \\
&= \sum_{m=1}^{M-1} (a_{11})_{mn} (\delta_x u_{m+\frac{1}{2},n} - \delta_x u_{m-\frac{1}{2},n}) u_{mn} \\
&= \sum_{m=2}^{M} (a_{11})_{m-1,n}\, \delta_x u_{m-\frac{1}{2},n}\, u_{m-1,n} - \sum_{m=1}^{M-1} (a_{11})_{mn}\, \delta_x u_{m-\frac{1}{2},n}\, u_{mn} \\
&= \sum_{m=1}^{M} (a_{11})_{m-1,n}\, \delta_x u_{m-\frac{1}{2},n}\, u_{m-1,n} - \sum_{m=1}^{M} (a_{11})_{mn}\, \delta_x u_{m-\frac{1}{2},n}\, u_{mn} \\
&= \sum_{m=1}^{M} (a_{11})_{mn}\, \delta_x u_{m-\frac{1}{2},n}\, (u_{m-1,n} - u_{mn}) \\
&\quad + \sum_{m=1}^{M} [(a_{11})_{m-1,n} - (a_{11})_{mn}]\, \delta_x u_{m-\frac{1}{2},n}\, u_{m-1,n} \\
&= -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} (\delta_x u_{m-\frac{1}{2},n})^2 - h_1 \sum_{m=0}^{M-1} (\delta_x a_{11})_{m+\frac{1}{2},n}\, \delta_x u_{m+\frac{1}{2},n}\, u_{mn},
\end{aligned}$$

we have

$$
\begin{aligned}
& h_1 \sum_{m=1}^{M-1} (a_{11})_{mn}\, \delta_x^2 u_{mn}\, u_{mn} \\
= & -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} \frac{(\delta_x u_{m-\frac{1}{2},n})^2 + (\delta_x u_{m+\frac{1}{2},n})^2}{2} \\
& -h_1 \sum_{m=0}^{M-1} (\delta_x a_{11})_{m+\frac{1}{2},n}\, \delta_x u_{m+\frac{1}{2},n}\, u_{m+\frac{1}{2},n} \\
= & -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} \frac{(\delta_x u_{m-\frac{1}{2},n})^2 + (\delta_x u_{m+\frac{1}{2},n})^2}{2} \\
& -\frac{1}{2} \sum_{m=0}^{M-1} (\delta_x a_{11})_{m+\frac{1}{2},n} \left(u_{m+1,n}^2 - u_{m,n}^2\right) \\
= & -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} \frac{(\delta_x u_{m-\frac{1}{2},n})^2 + (\delta_x u_{m+\frac{1}{2},n})^2}{2} \\
& +\frac{1}{2} \Bigg[\sum_{m=1}^{M-1} \left((\delta_x a_{11})_{m+\frac{1}{2},n} - (\delta_x a_{11})_{m-\frac{1}{2},n} \right) u_{mn}^2 \\
& \qquad + (\delta_x a_{11})_{\frac{1}{2},n} u_{0n}^2 - (\delta_x a_{11})_{M-\frac{1}{2},n} u_{Mn}^2 \Bigg] \\
\le & -h_1 \sum_{m=1}^{M-1} (a_{11})_{mn} \frac{(\delta_x u_{m-\frac{1}{2},n})^2 + (\delta_x u_{m+\frac{1}{2},n})^2}{2} \\
& +\frac{1}{2} c_1 h_1 \left(\frac{1}{2} u_{0n}^2 + \sum_{m=1}^{M-1} u_{mn}^2 + \frac{1}{2} u_{Mn}^2 \right).
\end{aligned}
$$

Here we have used the relations

$$
\left| (\delta_x a_{11})_{m+\frac{1}{2},n} - (\delta_x a_{11})_{m-\frac{1}{2},n} \right| \le c_1 h_1,
$$

$$
|(\delta_x a_{11})_{\frac{1}{2},n}| \le \frac{1}{2} c_1 h_1, \quad |(\delta_x a_{11})_{M-\frac{1}{2},n}| \le \frac{1}{2} c_1 h_1,
$$

which hold because of

$$
c_1 = \max_{(x,y,\tau)\in\Omega\times[0,T]} \left| \frac{\partial^2 a_{11}(x,y,\tau)}{\partial x^2} \right|_t \quad \text{and} \quad \frac{\partial a_{11}(x,y,\tau)}{\partial x} \Big|_{x=x_l \text{ or } x_u} = 0.
$$

Inserting the above equality into the equality (7.51), we obtain the inequality (7.49).

It is clear that for the second inequality in Proposition 7.1, the proof is almost the same as the proof for the first one. The concrete proof is omitted here. ∎

Proposition 7.2

$$\begin{aligned}&B_2 \equiv (a_{12}\varDelta_x\varDelta_y u, u)\\
\le& -\frac{1}{4}h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\Big[\delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} + \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}}\\
&+\delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} + \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}}\Big] + \frac{1}{2}c_2\|u\|^2. \qquad (7.52)\end{aligned}$$

Proof. Because $a_{12} = 0$ on all the boundary points, the expression of $(a_{12}\varDelta_x\varDelta_y u, u)$ can be written as follows:

$$\begin{aligned}B_2 &= h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}(\varDelta_x\varDelta_y u)_{mn}u_{mn}\\
&= \frac{1}{4}h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\Big(\delta_x\delta_y u_{m-\frac{1}{2},n-\frac{1}{2}} + \delta_x\delta_y u_{m+\frac{1}{2},n-\frac{1}{2}}\\
&\qquad\qquad +\delta_x\delta_y u_{m-\frac{1}{2},n+\frac{1}{2}} + \delta_x\delta_y u_{m+\frac{1}{2},n+\frac{1}{2}}\Big)u_{mn}\\
&= \frac{1}{4}\Big[h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m-\frac{1}{2},n-\frac{1}{2}}\ u_{mn}\\
&\quad +h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m+\frac{1}{2},n-\frac{1}{2}}\ u_{mn}\\
&\quad +h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m-\frac{1}{2},n+\frac{1}{2}}\ u_{mn}\\
&\quad +h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m+\frac{1}{2},n+\frac{1}{2}}\ u_{mn}\Big]\\
&\equiv \frac{1}{4}(B_{21}+B_{22}+B_{23}+B_{24}). \qquad (7.53)\end{aligned}$$

For B_{21}, we have

$$\begin{aligned}B_{21} &= h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m-\frac{1}{2},n-\frac{1}{2}}\ u_{mn}\\
&= h_2\sum_{n=1}^{N-1}\sum_{m=1}^{M-1}(a_{12})_{mn}(\delta_y u_{m,n-\frac{1}{2}} - \delta_y u_{m-1,n-\frac{1}{2}})u_{mn}\\
&= h_2\sum_{n=1}^{N-1}\left[\sum_{m=1}^{M-1}(a_{12})_{mn}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{mn} - \sum_{m=0}^{M-2}(a_{12})_{m+1,n}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m+1,n}\right]\\
&= h_2\sum_{n=1}^{N-1}\left[\sum_{m=0}^{M-1}(a_{12})_{mn}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{mn} - \sum_{m=0}^{M-1}(a_{12})_{m+1,n}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m+1,n}\right]\end{aligned}$$

$$
\begin{aligned}
&= h_2 \sum_{n=1}^{N-1} \Big[\sum_{m=0}^{M-1} (a_{12})_{mn}\ \delta_y u_{m,n-\frac{1}{2}}\ (u_{mn} - u_{m+1,n}) \\
&\qquad + \sum_{m=0}^{M-1} \left[(a_{12})_{mn} - (a_{12})_{m+1,n}\right]\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m+1,n} \Big] \\
&= -h_1 h_2 \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} (a_{12})_{mn}\ \delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \\
&\quad -h_1 h_2 \sum_{m=0}^{M-1} \sum_{n=1}^{N-1} (\delta_x a_{12})_{m+\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m+1,n}\ . \qquad (7.54)
\end{aligned}
$$

For B_{22}, we have

$$
\begin{aligned}
B_{22} &= h_1 h_2 \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} (a_{12})_{mn}\ \delta_x \delta_y u_{m+\frac{1}{2},n-\frac{1}{2}}\ u_{mn} \\
&= h_2 \sum_{n=1}^{N-1} \sum_{m=1}^{M-1} (a_{12})_{mn} (\delta_y u_{m+1,n-\frac{1}{2}} - \delta_y u_{m,n-\frac{1}{2}}) u_{mn} \\
&= h_2 \sum_{n=1}^{N-1} \left[\sum_{m=0}^{M-1} (a_{12})_{mn}\ \delta_y u_{m+1,n-\frac{1}{2}}\ u_{mn} \right. \\
&\qquad \left. - \sum_{m=1}^{M} (a_{12})_{mn}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m,n} \right] \\
&= h_2 \sum_{n=1}^{N-1} \left[\sum_{m=1}^{M} \left[(a_{12})_{m-1,n} - (a_{12})_{m,n}\right]\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m-1,n} \right. \\
&\qquad \left. - \sum_{m=1}^{M-1} (a_{12})_{mn}\ \delta_y u_{m,n-\frac{1}{2}}\ (u_{m,n} - u_{m-1,n}) \right] \\
&= h_2 \sum_{n=1}^{N-1} \left[-h_1 \sum_{m=1}^{M} (\delta_x a_{12})_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}}\ u_{m-1,n} \right. \\
&\qquad \left. -h_1 \sum_{m=1}^{M-1} (a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \right] \\
&= -h_1 h_2 \sum_{n=1}^{N-1} \left[\sum_{m=1}^{M-1} (a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \right. \\
&\qquad \left. + \sum_{m=0}^{M-1} (\delta_x a_{12})_{m+\frac{1}{2},n}\ \delta_y u_{m+1,n-\frac{1}{2}}\ u_{m,n} \right]
\end{aligned}
$$

$$\begin{aligned}
&= -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \\
&\quad -h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\ \delta_y u_{m+1,n-\frac{1}{2}}\ u_{m,n}\ .
\end{aligned} \tag{7.55}$$

We can see that during deriving the equalities (7.54) and (7.55), the subscripts n and $n-\frac{1}{2}$ are unchanged. Thus, from the equalities (7.54) and (7.55), for B_{23} and B_{24}, we can have

$$\begin{aligned}
B_{23} &= -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
&\quad -h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}}\ u_{m+1,n}\ ;
\end{aligned} \tag{7.56}$$

$$\begin{aligned}
B_{24} &= -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
&\quad -h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\ \delta_y u_{m+1,n+\frac{1}{2}}\ u_{mn}\ .
\end{aligned} \tag{7.57}$$

Putting the second terms in the last expressions of B_{21}, B_{22}, B_{23}, and B_{24} in the expressions (7.54)–(7.57) together yields

$$\begin{aligned}
&-h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(\delta_y u_{m,n-\frac{1}{2}} + \delta_y u_{m,n+\frac{1}{2}})u_{m+1,n} \\
&-h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(\delta_y u_{m+1,n-\frac{1}{2}} + \delta_y u_{m+1,n+\frac{1}{2}})u_{mn} \\
= &-h_1 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,[(u_{m,n+1} - u_{m,n-1})u_{m+1,n} \\
&\qquad\qquad +(u_{m+1,n+1} - u_{m+1,n-1})u_{mn}] \\
= &-h_1 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(u_{m+1,n+1}u_{mn} + u_{m,n+1}u_{m+1,n} \\
&\qquad\qquad -u_{m+1,n-1}u_{mn} - u_{m,n-1}u_{m+1,n}) \\
= &-h_1 \sum_{m=0}^{M-1}\Bigg[\sum_{n=0}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(u_{m+1,n+1}u_{mn} + u_{m,n+1}u_{m+1,n}) \\
&\qquad -\sum_{n=0}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n+1}\,(u_{m+1,n}u_{m,n+1} + u_{mn}u_{m+1,n+1})\Bigg]
\end{aligned}$$

$$\begin{aligned}
&= h_1h_2\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(\delta_y\delta_x a_{12})_{m+\frac{1}{2},n+\frac{1}{2}}\left(u_{m+1,n+1}u_{mn}+u_{m,n+1}u_{m+1,n}\right)\\
&\le \frac{1}{2}c_2h_1h_2\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left(u_{m+1,n+1}^2+u_{mn}^2+u_{m,n+1}^2+u_{m+1,n}^2\right)\\
&= 2c_2\|u\|^2. \qquad (7.58)
\end{aligned}$$

Here we have used $(\delta_x a_{12})_{m+\frac{1}{2},0}=(\delta_x a_{12})_{m+\frac{1}{2},N}=0$ and another form of the definition of inner product (7.42).

Thus, inserting the equalities (7.54)–(7.57) into the expression (7.53) and using the inequality (7.58), we get

$$\begin{aligned}
B_2 = &-\frac{1}{4}h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\Big(\delta_x u_{m+\frac{1}{2},n}\,\delta_y u_{m,n-\frac{1}{2}}+\delta_x u_{m-\frac{1}{2},n}\,\delta_y u_{m,n-\frac{1}{2}}\\
&\qquad +\delta_x u_{m+\frac{1}{2},n}\,\delta_y u_{m,n+\frac{1}{2}}+\delta_x u_{m-\frac{1}{2},n}\,\delta_y u_{m,n+\frac{1}{2}}\Big)\\
&-\frac{1}{4}h_1h_2\sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(\delta_y u_{m,n-\frac{1}{2}}+\delta_y u_{m,n+\frac{1}{2}})u_{m+1,n}\\
&-\frac{1}{4}h_1h_2\sum_{m=0}^{M-1}\sum_{n=1}^{N-1}(\delta_x a_{12})_{m+\frac{1}{2},n}\,(\delta_y u_{m+1,n-\frac{1}{2}}+\delta_y u_{m+1,n+\frac{1}{2}})u_{mn}\\
\le &-\frac{1}{4}h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\Big(\delta_x u_{m+\frac{1}{2},n}\,\delta_y u_{m,n-\frac{1}{2}}+\delta_x u_{m-\frac{1}{2},n}\,\delta_y u_{m,n-\frac{1}{2}}\\
&+\delta_x u_{m+\frac{1}{2},n}\,\delta_y u_{m,n+\frac{1}{2}}+\delta_x u_{m-\frac{1}{2},n}\,\delta_y u_{m,n+\frac{1}{2}}\Big)+\frac{1}{2}c_2\|u\|^2. \quad \blacksquare
\end{aligned}$$

Proposition 7.3 *For $(b_1\varDelta_x u,u)$ and $(b_2\varDelta_y u,u)$, we have*

$$B_4\equiv(b_1\varDelta_x u,u)\le\frac{1}{2}c_4\|u\|^2 \qquad (7.59)$$

and

$$B_5\equiv(b_2\varDelta_y u,u)\le\frac{1}{2}c_5\|u\|^2. \qquad (7.60)$$

Proof. Because $(b_1)_{0,n}=(b_1)_{M,n}$ for $n=0,1,\cdots,N$, the concrete expression for $(b_1\varDelta_x u,u)$ is

$$\begin{aligned}
B_4 = h_1h_2\Bigg[&\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(b_1)_{mn}\,\varDelta_x u_{mn}\,u_{mn}+\frac{1}{2}\sum_{m=1}^{M-1}(b_1)_{m0}\,\varDelta_x u_{m0}\,u_{m0}\\
&+\frac{1}{2}\sum_{m=1}^{M-1}(b_1)_{mN}\,\varDelta_x u_{mN}\,u_{mN}\Bigg].
\end{aligned}$$

For any n, we have

$$\begin{aligned}
& h_1 \sum_{m=1}^{M-1} (b_1)_{mn}\ \varDelta_x u_{mn}\ u_{mn} \\
= & \frac{1}{2} \sum_{m=1}^{M-1} (b_1)_{mn} (u_{m+1,n} - u_{m-1,n}) u_{mn} \\
= & \frac{1}{2} \left(\sum_{m=1}^{M-1} (b_1)_{mn} u_{mn} u_{m+1,n} - \sum_{m=0}^{M-2} (b_1)_{m+1,n} u_{mn} u_{m+1,n} \right) \\
= & -\frac{1}{2} h_1 \sum_{m=0}^{M-1} (\delta_x b_1)_{m+\frac{1}{2},n}\ u_{mn} u_{m+1,n} \\
\le & \frac{1}{2} c_4 h_1 \Big(\frac{1}{2}\ u_{0n}^2 + \sum_{m=1}^{M-1} u_{mn}^2 + \frac{1}{2}\ u_{Mn}^2 \Big).
\end{aligned}$$

Adding them together yields

$$B_4 \le \frac{1}{2} c_4 \|u\|^2.$$

It is easy to see that changing x to y and m to n during the derivation above, we can prove the second inequality in Proposition 7.3. Thus, we have proved the conclusion we need. ■

Proposition 7.4

$$B_6 \equiv (cu, u) \le c_6 \|u\|^2. \tag{7.61}$$

Proof. Since $|c_{mn}^k| \le c_6$, it is easy to see the validity of the inequality (7.61). ■

The proof of Lemma 7.1 Based on these inequalities and noticing the matrix

$$\begin{pmatrix} a_{11}(x,y,\tau) & a_{12}(x,y,\tau) \\ a_{12}(x,y,\tau) & a_{22}(x,y,\tau) \end{pmatrix}$$

is semi-positive, we can prove the lemma immediately. Adding the relations (7.49), (7.52), (7.50), (7.59), (7.60), and (7.61), then using the inequality (7.37), we get

$$\begin{aligned}
& B_1 + 2B_2 + B_3 + B_4 + B_5 + B_6 \\
\le & \frac{1}{2} (c_1 + 2c_2 + c_3 + c_4 + c_5 + 2c_6) \|u\|^2 \\
& -\frac{1}{4} h_1 h_2 \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} \Big\{ (a_{11})_{mn} \Big[2(\delta_x u_{m-\frac{1}{2},n})^2 + 2(\delta_x u_{m+\frac{1}{2},n})^2 \Big]
\end{aligned}$$

$$
\begin{aligned}
&+2(a_{12})_{mn}\Big[\delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} + \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \\
&+\delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} + \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}}\Big] \\
&+(a_{22})_{mn}\Big[2(\delta_y u_{m,n-\frac{1}{2}})^2 + 2(\delta_y u_{m,n+\frac{1}{2}})^2\Big]\Big\} \\
= &\frac{c}{2}\|u\|^2 - \frac{1}{4}h_1h_2\sum_{m=1}^{M-1}\sum_{n=1}^{N-1}\Big\{ \\
&\Big[(a_{11})_{mn}(\delta_x u_{m+\frac{1}{2},n})^2 + 2(a_{12})_{mn}\ \delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \\
&\qquad\qquad +(a_{22})_{mn}(\delta_y u_{m,n-\frac{1}{2}})^2\Big] \\
&+\Big[(a_{11})_{mn}(\delta_x u_{m-\frac{1}{2},n})^2 + 2(a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n-\frac{1}{2}} \\
&\qquad\qquad +(a_{22})_{mn}(\delta_y u_{m,n-\frac{1}{2}})^2\Big] \\
&+\Big[(a_{11})_{mn}(\delta_x u_{m+\frac{1}{2},n})^2 + 2(a_{12})_{mn}\ \delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
&\qquad\qquad +(a_{22})_{mn}(\delta_y u_{m,n+\frac{1}{2}})^2\Big] \\
&+\Big[(a_{11})_{mn}(\delta_x u_{m-\frac{1}{2},n})^2 + 2(a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
&\qquad\qquad +(a_{22})_{mn}(\delta_y u_{m,n+\frac{1}{2}})^2\Big]\Big\} \\
\le &\frac{c}{2}\|u\|^2.
\end{aligned}
$$

This completes the proof of Lemma 7.1. □

7.4.3 ‡Solvability and Stability

In this subsection, we will prove the solvability and stability of the two-dimensional finite-difference scheme (7.46)–(7.47).

Theorem 7.1 *If $\Delta\tau < 1/c$, then the difference scheme (7.46)–(7.47) is uniquely solvable.*

Proof. Suppose $\{u_{mn}^k \mid 0 \le m \le M, 0 \le n \le N\}$ has been determined. Then the difference scheme (7.46) is a linear system about $\{u_{mn}^{k+1} \mid 0 \le m \le M, 0 \le n \le N\}$. Consider its homogeneous system

$$
\begin{aligned}
\frac{1}{\Delta\tau}u_{mn}^{k+1} = &\frac{1}{2}(a_{11})_{mn}^{k+\frac{1}{2}}\delta_x^2 u_{mn}^{k+1} + (a_{12})_{mn}^{k+\frac{1}{2}}\Delta_x\Delta_y u_{mn}^{k+1} + \frac{1}{2}(a_{22})_{mn}^{k+\frac{1}{2}}\delta_y^2 u_{mn}^{k+1} \\
&+\frac{1}{2}(b_1)_{mn}^{k+\frac{1}{2}}\Delta_x u_{mn}^{k+1} + \frac{1}{2}(b_2)_{mn}^{k+\frac{1}{2}}\Delta_y u_{mn}^{k+1} + \frac{1}{2}c_{mn}^{k+\frac{1}{2}}u_{mn}^{k+1}, \\
&\qquad 0 \le m \le M, \quad 0 \le n \le N. \qquad (7.62)
\end{aligned}
$$

Taking the inner product of equality (7.62) with $2u^{k+1}$ and using Lemma 7.1, we have

$$\begin{aligned}\frac{2}{\Delta\tau}\|u^{k+1}\|^2 &= \left((a_{11})^{k+\frac{1}{2}}\delta_x^2 u^{k+1}, u^{k+1}\right) + 2\left((a_{12})^{k+\frac{1}{2}}\Delta_x\Delta_y u^{k+1}, u^{k+1}\right)\\ &\quad+\left((a_{22})^{k+\frac{1}{2}}\delta_y^2 u^{k+1}, u^{k+1}\right) + \left((b_1)^{k+\frac{1}{2}}\Delta_x u^{k+1}, u^{k+1}\right)\\ &\quad+\left((b_2)^{k+\frac{1}{2}}\Delta_y u^{k+1}, u^{k+1}\right) + \left(c^{k+\frac{1}{2}} u^{k+1}, u^{k+1}\right)\\ &\le \frac{c}{2}\|u^{k+1}\|^2. \end{aligned} \tag{7.63}$$

If $\Delta\tau < 1/c$, then $\|u^{k+1}\| = 0$. This completes the proof. ■

Theorem 7.2 *If $\Delta\tau \le 2/[3(1+c)]$, then the solution to the difference scheme (7.46)–(7.47) satisfies*

$$\|u^{k+1}\|^2 \le \mathrm{e}^{3(c+1)T/2}\left(\|u^0\|^2 + \frac{3}{2}\Delta\tau\sum_{l=0}^{k}\|g^{l+\frac{1}{2}}\|^2\right), \quad 0 \le k \le K-1. \tag{7.64}$$

Proof. Taking the inner product of Eq. (7.46) with $u^{k+\frac{1}{2}}$ and using Lemma 7.1, we have

$$\begin{aligned}&\frac{1}{2\Delta\tau}\left(\|u^{k+1}\|^2 - \|u^k\|^2\right)\\ &= \left((a_{11})^{k+\frac{1}{2}}\delta_x^2 u^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right) + 2\left((a_{12})^{k+\frac{1}{2}}\Delta_x\Delta_y u^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right)\\ &\quad+\left((a_{22})^{k+\frac{1}{2}}\delta_y^2 u^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right) + \left((b_1)^{k+\frac{1}{2}}\Delta_x u^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right)\\ &\quad+\left((b_2)^{k+\frac{1}{2}}\Delta_y u^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right) + \left(c^{k+\frac{1}{2}} u^{k+1}, u^{k+\frac{1}{2}}\right) + \left(g^{k+\frac{1}{2}}, u^{k+\frac{1}{2}}\right)\\ &\le \frac{c}{2}\|u^{k+\frac{1}{2}}\|^2 + \frac{1}{2}\|g^{k+\frac{1}{2}}\|^2 + \frac{1}{2}\|u^{k+\frac{1}{2}}\|^2, \quad 0 \le k \le K-1,\end{aligned}$$

from which we further obtain

$$\begin{aligned}\|u^{k+1}\|^2 &\le \|u^k\|^2 + (1+c)\Delta\tau\|u^{k+\frac{1}{2}}\|^2 + \Delta\tau\|g^{k+\frac{1}{2}}\|^2\\ &\le \|u^k\|^2 + \frac{1+c}{2}\Delta\tau\left(\|u^k\|^2 + \|u^{k+1}\|^2\right) + \Delta\tau\|g^{k+\frac{1}{2}}\|^2,\\ &\qquad 0 \le k \le K-1.\end{aligned}$$

If $1 - \dfrac{1+c}{2}\Delta\tau > 0$, then the inequality can be rewritten as

$$\|u^{k+1}\|^2 \le \frac{1+\frac{1+c}{2}\Delta\tau}{1-\frac{1+c}{2}\Delta\tau}\|u^k\|^2 + \frac{\Delta\tau}{1-\frac{1+c}{2}\Delta\tau}\|g^{k+\frac{1}{2}}\|^2.$$

It is clear that for $\bar{C} > 2$, when $\Delta\tau$ is small enough, we can have $\dfrac{1+\frac{1+c}{2}\Delta\tau}{1-\frac{1+c}{2}\Delta\tau} \le 1+\bar{C}\frac{1+c}{2}\Delta\tau$. Let us take $\bar{C} = 3$; then we can easily find that the corresponding condition for $\Delta\tau$ is $\Delta\tau \le 2/[3(c+1)]$ and that in this case $1-\frac{1+c}{2}\Delta\tau \ge \frac{2}{3}$. Thus, when $\Delta\tau \le 2/[3(c+1)]$, we have

$$\|u^{k+1}\|^2 \le \left(1+\frac{3(c+1)}{2}\Delta\tau\right)\|u^k\|^2 + \frac{3}{2}\Delta\tau\|g^{k+\frac{1}{2}}\|^2, \quad 0 \le k \le K-1.$$

From this discrete Gronwall inequality, we finally arrive at

$$\|u^{k+1}\|^2 \le \mathrm{e}^{3(c+1)T/2}\left[\|u^0\|^2 + \frac{3}{2}\Delta\tau\sum_{l=0}^{k}\|g^{l+\frac{1}{2}}\|^2\right], \quad 0 \le k \le K-1.$$

This completes the proof. ■

The method used here to prove the stability is usually called the energy method for stability analysis.

7.4.4 ‡Convergence

For the convergence of the finite-difference scheme (7.46)–(7.47), we have

Theorem 7.3 *Let $\{U_{mn}^k\}$ be the solution of the problem (7.32)–(7.33) and $\{u_{mn}^k\}$ be the solution of Eqs. (7.46)–(7.47). Denote*

$$e_{mn}^k = U_{mn}^k - u_{mn}^k, \quad 0 \le m \le M, \quad 0 \le n \le N, \quad 0 \le k \le K.$$

If $\Delta\tau \le 2/[3(c+1)]$, then we have

$$\|e^{k+1}\| \le \mathrm{e}^{3(c+1)T/4}\sqrt{\frac{3(x_u-x_l)(y_u-y_l)T}{2}}\,c_0\left(h_1^2+h_2^2+\Delta\tau^2\right),$$
$$0 \le k \le K-1.$$

Proof. Subtracting the equalities (7.46) and (7.47) from the equalities (7.43) and (7.45), respectively, we obtain the error equations

$$\begin{aligned}\frac{1}{\Delta\tau}(e_{mn}^{k+1}-e_{mn}^k) = {} & (a_{11})_{mn}^{k+\frac{1}{2}}\delta_x^2 e_{mn}^{k+\frac{1}{2}} + 2(a_{12})_{mn}^{k+\frac{1}{2}}\Delta_x\Delta_y e_{mn}^{k+\frac{1}{2}} \\ & +(a_{22})_{mn}^{k+\frac{1}{2}}\delta_y^2 e_{mn}^{k+\frac{1}{2}} + (b_1)_{mn}^{k+\frac{1}{2}}\Delta_x e_{mn}^{k+\frac{1}{2}} \\ & +(b_2)_{mn}^{k+\frac{1}{2}}\Delta_y e_{mn}^{k+\frac{1}{2}} + c_{mn}^{k+\frac{1}{2}} e_{mn}^{k+\frac{1}{2}} + R_{mn}^{k+\frac{1}{2}}, \\ & 0 \le m \le M, \quad 0 \le n \le N, \quad 0 \le k \le K-1, \end{aligned} \quad (7.65)$$

$$e_{mn}^0 = 0, \quad 0 \le m \le M, \quad 0 \le n \le N. \quad (7.66)$$

Taking the inner product of the system (7.65) with $e^{k+\frac{1}{2}}$ and using Lemma 7.1, we have

$$\begin{aligned}
&\frac{1}{2\varDelta\tau}\left(\|e^{k+1}\|^2-\|e^k\|^2\right)\\
&=\left((a_{11})^{k+\frac{1}{2}}\delta_x^2 e^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)+2\left((a_{12})^{k+\frac{1}{2}}\varDelta_x\varDelta_y e^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)\\
&\quad+\left((a_{22})^{k+\frac{1}{2}}\delta_y^2 e^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)+\left((b_1)^{k+\frac{1}{2}}\varDelta_x e^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)\\
&\quad+\left((b_2)^{k+\frac{1}{2}}\varDelta_y e^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)+\left(c^{k+\frac{1}{2}}e^{k+1},e^{k+\frac{1}{2}}\right)+\left(R^{k+\frac{1}{2}},e^{k+\frac{1}{2}}\right)\\
&\le\frac{c}{2}\|e^{k+\frac{1}{2}}\|^2+\frac{1}{2}\|R^{k+\frac{1}{2}}\|^2+\frac{1}{2}\|e^{k+\frac{1}{2}}\|^2,\quad 0\le k\le K-1,
\end{aligned}$$

from which we further get

$$\begin{aligned}
\|e^{k+1}\|^2&\le\|e^k\|^2+(1+c)\varDelta\tau\|e^{k+\frac{1}{2}}\|^2+\varDelta\tau\|R^{k+\frac{1}{2}}\|^2\\
&\le\|e^k\|^2+\frac{1+c}{2}\varDelta\tau\left(\|e^k\|^2+\|e^{k+1}\|^2\right)+\varDelta\tau\|R^{k+\frac{1}{2}}\|^2,\\
&\qquad 0\le k\le K-1.
\end{aligned}$$

Using the condition (7.44) and when $\varDelta\tau\le 2/[3(c+1)]$, we can rewrite this inequality as

$$\begin{aligned}
\|e^{k+1}\|^2&\le\left(1+\frac{3(c+1)}{2}\varDelta\tau\right)\|e^k\|^2+\frac{3}{2}\varDelta\tau\|R^{k+\frac{1}{2}}\|^2\\
&\le\left(1+\frac{3(c+1)}{2}\varDelta\tau\right)\|e^k\|^2\\
&\quad+\frac{3}{2}(x_u-x_l)(y_u-y_l)c_0^2\varDelta\tau\left(h_1^2+h_2^2+\varDelta\tau^2\right)^2,\\
&\qquad 0\le k\le K-1.
\end{aligned}$$

The Gronwall inequality gives

$$\|e^{k+1}\|^2\le \mathrm{e}^{3(c+1)T/2}\frac{3(x_u-x_l)(y_u-y_l)T}{2}c_0^2\left(h_1^2+h_2^2+\varDelta\tau^2\right)^2,\ 0\le k\le K-1,$$

or

$$\|e^{k+1}\|\le \mathrm{e}^{3(c+1)T/4}\sqrt{3\frac{(x_u-x_l)(y_u-y_l)T}{2}}\,c_0\left(h_1^2+h_2^2+\varDelta\tau^2\right),$$
$$0\le k\le K-1.$$

This completes the proof. ■

For the solution to the difference scheme (7.46)–(7.47), we can also use the extrapolation technique to improve the accuracy of the numerical solutions when solutions are smooth. The idea is the same as what is described in Sect. 7.3. Based on the results given in this subsection, some theoretical conclusions on the extrapolation technique can be obtained. For details, see the paper [78] by Sun and Zhu.

Problems

Table 7.1. Problems and Sections

Problems	Sections	Problems	Sections	Problems	Sections
1–5	7.1	6–15	7.2	16–18	7.3
19–21	7.4				

1. *Let f_m^n denote $f(m\Delta x, n\Delta \tau)$. Find the truncation error of the explicit difference scheme

$$\begin{aligned}\frac{u_m^{n+1}-u_m^n}{\Delta\tau} &= a_m^n \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2} \\ &\quad + b_m^n \frac{u_{m+1}^n - u_{m-1}^n}{2\Delta x} + c_m^n u_m^n\end{aligned}$$

to the parabolic partial differential equation

$$\frac{\partial u}{\partial \tau} = a(x,\tau)\frac{\partial^2 u}{\partial x^2} + b(x,\tau)\frac{\partial u}{\partial x} + c(x,\tau)u.$$

2. Show that the truncation error of the Crank–Nicolson scheme for the heat equation at the point $(x_m, \tau^{n+1/2})$ is in the following form:

$$\Delta\tau^2\left[\frac{1}{24}\frac{\partial^3 u}{\partial \tau^3}(x_m,\eta^{(1)}) - \frac{a}{8}\frac{\partial^4 u}{\partial x^2 \partial \tau^2}(x_m,\eta^{(2)})\right] - \frac{\Delta x^2 a}{12}\frac{\partial^4 u}{\partial x^4}(\xi,\eta^{(3)}),$$

where $\xi \in (x_{m-1}, x_{m+1})$, $\eta^{(k)} \in (\tau^n, \tau^{n+1})$, $k = 1, 2, 3$, and a is the conductivity coefficient in the heat equation.

3. *Let f_m^n denote $f(m\Delta x, n\Delta \tau)$. Find the truncation error of the implicit difference scheme

$$\begin{aligned}\frac{u_m^{n+1}-u_m^n}{\Delta\tau} &= \frac{a_m^{n+1/2}}{2}\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2} + \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2}\right) \\ &\quad + \frac{b_m^{n+1/2}}{2}\left(\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2\Delta x} + \frac{u_{m+1}^n - u_{m-1}^n}{2\Delta x}\right) \\ &\quad + \frac{c_m^{n+1/2}}{2}(u_m^{n+1} + u_m^n)\end{aligned}$$

to the parabolic partial differential equation

$$\frac{\partial u}{\partial \tau} = a(x,\tau)\frac{\partial^2 u}{\partial x^2} + b(x,\tau)\frac{\partial u}{\partial x} + c(x,\tau)u.$$

4. The heat equation

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}$$

can also be discretized by

$$\frac{u_m^{n+1} - u_m^n}{\Delta\tau} = \theta \left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2} \right) + (1-\theta) \left(\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2} \right)$$

or

$$u_m^{n+1} - \theta\alpha(u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}) = u_m^n + (1-\theta)\alpha(u_{m+1}^n - 2u_m^n + u_{m-1}^n),$$

where $0 \le \theta \le 1$ and $\alpha = \Delta\tau/\Delta x^2$. This scheme is called the θ–scheme. It is clear that when $\theta = 0$, the scheme reduces to the explicit scheme and when $\theta = 1/2$, the scheme becomes the Crank–Nicolson scheme. Show that the order of truncation error of the θ–scheme is

$$O\left((1-2\theta)\,\Delta\tau + \Delta\tau^2 + \Delta x^2\right).$$

(Hint: Discretize the partial differential equation at $x = x_m$ and $\tau = \tau^{n+\theta}$.)

5. Consider the parabolic partial differential equation

$$\frac{\partial u}{\partial \tau} = a(x,\tau)\frac{\partial^2 u}{\partial x^2} + b(x,\tau)\frac{\partial u}{\partial x} + c(x,\tau)u,$$

which is defined for $x \in [0,1]$ and $\tau \ge 0$. Here $a(x,\tau) \ge 0$ holds and we suppose that $\frac{\partial a}{\partial x}$ is bounded. Assuming that $u(x,\tau)$ is given, we want to determine $u(x,\tau+\Delta\tau)$ with $\Delta\tau > 0$ for $x \in [0,1]$.
   (a) Under what conditions on $a(x,\tau)$ and $b(x,\tau)$ a boundary condition is needed and under what conditions no boundary condition is needed at $x = 0$ and $x = 1$?
   (b) Suppose that an explicit scheme will be used. How do we determine $u(0,\tau+\Delta\tau)$ and $u(1,\tau+\Delta\tau)$ if no boundary condition should be given?

6. *Consider the three-point explicit finite-difference scheme:

$$u_m^{n+1} = a_m u_{m-1}^n + b_m u_m^n + c_m u_{m+1}^n, \quad m = 1, 2, \cdots, M-1,$$

where $a_m \ge 0$, $b_m = 1 - a_m - c_m \ge 0$, $c_m \ge 0$ and $a_0 = c_M = 0$. Show

$$\max_{1\le m\le M-1} |u_m^{n+1}| \le \max_{1\le m\le M-1} |u_m^n|.$$

This means that the numerical procedure is stable under the maximum norm.

7. Consider the equation

$$\lambda \mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x} \quad \text{or} \quad \mathbf{A}^{-1}\mathbf{B}\mathbf{x} = \lambda \mathbf{x},$$

where $\mathbf{A}$ and $\mathbf{B}$ are $(M-1) \times (M-1)$ matrices and their concrete expressions are

$$\mathbf{A} = \begin{bmatrix} a_0 & a_1 & 0 & \cdots & 0 \\ a_1 & a_0 & a_1 & \ddots & \vdots \\ 0 & a_1 & a_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_1 \\ 0 & \cdots & 0 & a_1 & a_0 \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} b_0 & b_1 & 0 & \cdots & 0 \\ b_1 & b_0 & b_1 & \ddots & \vdots \\ 0 & b_1 & b_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_1 \\ 0 & \cdots & 0 & b_1 & b_0 \end{bmatrix}.$$

Find $M-1$ linearly independent eigenvectors of $\mathbf{A}^{-1}\mathbf{B}$ and their associated eigenvalues.

8. Consider the equation

$$\lambda \mathbf{A}_2\mathbf{x} = \mathbf{B}_2\mathbf{x}$$

or

$$\mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{x} = \lambda \mathbf{x},$$

where $\mathbf{A}_2$ and $\mathbf{B}_2$ are $M \times M$ matrices and their concrete expressions are

$$\mathbf{A}_2 = \begin{bmatrix} a_0 & a_1 & 0 & \cdots & a_{-1} \\ a_{-1} & a_0 & a_1 & \ddots & \vdots \\ 0 & a_{-1} & a_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_1 \\ a_1 & \cdots & 0 & a_{-1} & a_0 \end{bmatrix}$$

and

$$\mathbf{B}_2 = \begin{bmatrix} b_0 & b_1 & 0 & \cdots & b_{-1} \\ b_{-1} & b_0 & b_1 & \ddots & \vdots \\ 0 & b_{-1} & b_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_1 \\ b_1 & \cdots & 0 & b_{-1} & b_0 \end{bmatrix}.$$

Find M linearly independent eigenvectors of $\mathbf{A}_2{}^{-1}\mathbf{B}_2$ and their associated eigenvalues.

9. (a) Consider an $M \times M$ matrix

$$\mathbf{A} = \begin{pmatrix} a & b & 0 & \cdots & \cdots & 0 & b \\ b & a & b & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & b & a & b \\ b & 0 & \cdots & \cdots & 0 & b & a \end{pmatrix}.$$

Suppose $a = q + 2/h^2$ and $b = -1/h^2$. Show that its eigenvalues are $\lambda_j = q + \dfrac{4}{h^2}\sin^2\dfrac{\theta_j}{2}$, $j = 0, 1, \cdots, M-1$, where $\theta_j = j\frac{2\pi}{M}$, and the corresponding eigenvectors are

$$\mathbf{v}_j = \begin{pmatrix} 1 \\ \cos\theta_j \\ \cos 2\theta_j \\ \vdots \\ \cos(M-1)\theta_j \end{pmatrix}, \quad j = 0, 1, \cdots, \text{int}\left(\frac{M}{2}\right),$$

and

$$\mathbf{v}_j = \begin{pmatrix} 0 \\ \sin\theta_j \\ \sin 2\theta_j \\ \vdots \\ \sin(M-1)\theta_j \end{pmatrix}, \quad j = \text{int}\left(\frac{M}{2}\right) + 1, \cdots, M-1,$$

respectively, where $\text{int}\left(\dfrac{M}{2}\right)$ is the integer part of $\dfrac{M}{2}$.

(b) Find the eigenvalues and eigenvectors of $\mathbf{A}^{-1}$.

(c) Suppose $a = \dfrac{q}{2} + \dfrac{2}{h^2}$ and $b = \dfrac{q}{4} - \dfrac{1}{h^2}$, find the eigenvalues and eigenvectors of $\mathbf{A}$ and $\mathbf{A}^{-1}$.

10. *Consider the explicit scheme

$$\frac{u_m^{n+1} - u_m^n}{\Delta\tau} = a\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2}, \quad m = 1, 2, \cdots, M-1$$

with $u_0^{n+1} = f_l(\tau^{n+1})$ and $u_M^{n+1} = f_u(\tau^{n+1})$. Determine when it is stable with respect to initial values in L_2 norm and when it is unstable. (Suppose $a > 0$.)

11. *Consider the implicit scheme

$$\frac{u_m^{n+1} - u_m^n}{\Delta\tau} = \frac{a}{2}\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta x^2} + \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2}\right),$$
$$m = 1, 2, \cdots, M-1$$

with $u_0^{n+1} = f_l(\tau^{n+1})$ and $u_M^{n+1} = f_u(\tau^{n+1})$. Show that it is always stable with respect to initial values in L_2 norm. (Suppose $a > 0$.)

12. By using the von Neumann method, show that for periodic problems, the θ–scheme for the heat equation

$$\begin{aligned} &u_m^{n+1} - \theta\alpha\left(u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}\right) \\ &= u_m^n + (1-\theta)\,\alpha\left(u_{m+1}^n - 2u_m^n + u_{m-1}^n\right) \end{aligned}$$

is stable for all $\alpha > 0$ if $\dfrac{1}{2} \le \theta \le 1$ and that it is stable for $0 < \alpha \le \dfrac{1}{2(1-2\theta)}$ if $0 < \theta < \dfrac{1}{2}$.

13. Consider the following parabolic partial differential equation:

$$\frac{\partial u}{\partial \tau} = a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2} + b_1\frac{\partial u}{\partial x} + b_2\frac{\partial u}{\partial y},$$

where $a_{11}(x, y, \tau) \ge 0, a_{22}(x, y, \tau) \ge 0, a_{12}(x, y, \tau) = \rho_{12}(x, y, \tau)\sqrt{a_{11}a_{22}}$ with $\rho_{12} \in [-1, 1]$, and b_1, b_2 are any functions of x, y, τ. This equation can be approximated by

(i)

$$\begin{aligned} &\frac{u_{m,n}^{k+1} - u_{m,n}^k}{\Delta\tau} \\ &= \frac{a_{11,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m+1,n}^{k+1} - 2u_{m,n}^{k+1} + u_{m-1,n}^{k+1}}{\Delta x^2} + \frac{u_{m+1,n}^k - 2u_{m,n}^k + u_{m-1,n}^k}{\Delta x^2}\right) \end{aligned}$$

$$
\begin{aligned}
&+a_{12,m,n}^{k+\frac{1}{2}}\left(\frac{u_{m+1,n+1}^{k+1}-u_{m+1,n-1}^{k+1}-u_{m-1,n+1}^{k+1}+u_{m-1,n-1}^{k+1}}{4\Delta x\Delta y}\right.\\
&\qquad\left.+\frac{u_{m+1,n+1}^{k}-u_{m+1,n-1}^{k}-u_{m-1,n+1}^{k}+u_{m-1,n-1}^{k}}{4\Delta x\Delta y}\right)\\
&+\frac{a_{22,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m,n+1}^{k+1}-2u_{m,n}^{k+1}+u_{m,n-1}^{k+1}}{\Delta y^2}+\frac{u_{m,n+1}^{k}-2u_{m,n}^{k}+u_{m,n-1}^{k}}{\Delta y^2}\right)\\
&+\frac{b_{1,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m+1,n}^{k+1}-u_{m-1,n}^{k+1}}{2\Delta x}+\frac{u_{m+1,n}^{k}-u_{m-1,n}^{k}}{2\Delta x}\right)\\
&+\frac{b_{2,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m,n+1}^{k+1}-u_{m,n-1}^{k+1}}{2\Delta y}+\frac{u_{m,n+1}^{k}-u_{m,n-1}^{k}}{2\Delta y}\right)\qquad\text{or}
\end{aligned}
$$

(ii)

$$
\begin{aligned}
&\frac{u_{m,n}^{k+1}-u_{m,n}^{k}}{\Delta\tau}\\
&=\frac{a_{11,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m+1,n}^{k+1}-2u_{m,n}^{k+1}+u_{m-1,n}^{k+1}}{\Delta x^2}+\frac{u_{m+1,n}^{k}-2u_{m,n}^{k}+u_{m-1,n}^{k}}{\Delta x^2}\right)\\
&+a_{12,m,n}^{k+\frac{1}{2}}\left(\frac{u_{m+1,n+1}^{k+1}-u_{m+1,n-1}^{k+1}-u_{m-1,n+1}^{k+1}+u_{m-1,n-1}^{k+1}}{4\Delta x\Delta y}\right.\\
&\qquad\left.+\frac{u_{m+1,n+1}^{k}-u_{m+1,n-1}^{k}-u_{m-1,n+1}^{k}+u_{m-1,n-1}^{k}}{4\Delta x\Delta y}\right)\\
&+\frac{a_{22,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{u_{m,n+1}^{k+1}-2u_{m,n}^{k+1}+u_{m,n-1}^{k+1}}{\Delta y^2}+\frac{u_{m,n+1}^{k}-2u_{m,n}^{k}+u_{m,n-1}^{k}}{\Delta y^2}\right)\\
&+\frac{b_{1,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{-u_{m+2,n}^{k+1}+4u_{m+1,n}^{k+1}-3u_{m,n}^{k+1}}{2\Delta x}\right.\\
&\qquad\left.+\frac{-u_{m+2,n}^{k}+4u_{m+1,n}^{k}-3u_{m,n}^{k}}{2\Delta x}\right)\\
&+\frac{b_{2,m,n}^{k+\frac{1}{2}}}{2}\left(\frac{3u_{m,n}^{k+1}-4u_{m,n-1}^{k+1}+u_{m,n-2}^{k+1}}{2\Delta y}+\frac{3u_{m,n}^{k}-4u_{m,n-1}^{k}+u_{m,n-2}^{k}}{2\Delta y}\right)
\end{aligned}
$$

if $b_1(x,y,\tau)\ge 0$ and $b_2(x,y,\tau)\le 0$. By the von Neumann method, show that they are stable.

(Hint:

(a) First show that the amplification factor λ can be written as $\lambda = \dfrac{1+a+ib}{1-a-ib}$.

(b) Then show that $|\lambda|^2 \le 1$ is equivalent to $|1-a-ib|^2 - |1+a+ib|^2 = -4a \ge 0$.

(c) Finally show $-4a \ge 0$ by using the following inequalities: (i) $A^2 + B^2 + 2\rho AB = (A+\rho B)^2 + B^2\left(1-\rho^2\right) \ge 0$ if $|\rho| \le 1$; (ii) $\cos 2\theta - 4\cos\theta + 3 = 2\left(\cos\theta - 1\right)^2 \ge 0$.)

14. *Show that if

$$\max_{0\le m\le M} \frac{x_m^2(1-x_m)^2\bar{\sigma}_m^2}{2}\frac{\varDelta\tau}{\varDelta x^2} \le \frac{1}{2},$$

then for the scheme with variable coefficients

$$\begin{aligned}\frac{u_m^{n+1}-u_m^n}{\varDelta\tau} &= \frac{1}{2}[x_m(1-x_m)\bar{\sigma}_m]^2\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\varDelta x^2}\\ &\quad + (r-D_0)x_m(1-x_m)\frac{u_{m+1}^n - u_{m-1}^n}{2\varDelta x}\\ &\quad - \left[r\left(1-x_m\right) + D_0x_m\right]u_m^n,\end{aligned}$$

the condition $|\lambda_\theta(x_m,\tau^n)| \le 1 + O(\varDelta\tau)$ is satisfied for any $x_m = m/M \in [0,1]$. (When you prove this result, you should derive the stability condition for explicit schemes by yourself.)

15. For the scheme with variable coefficients

$$\begin{aligned}&\frac{u_m^{n+1}-u_m^n}{\varDelta\tau}\\ =&\frac{1}{4}[x_m(1-x_m)\bar{\sigma}_m]^2\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\varDelta x^2} + \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\varDelta x^2}\right)\\ &+ \frac{1}{2}(r-D_0)x_m(1-x_m)\left(\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2\varDelta x} + \frac{u_{m+1}^n - u_{m-1}^n}{2\varDelta x}\right)\\ &- \frac{1}{2}\left[r\left(1-x_m\right) + D_0x_m\right]\left(u_m^{n+1} + u_m^n\right),\end{aligned}$$

show that the condition $|\lambda_\theta(x_m,\tau^n)| \le 1 + O(\varDelta\tau)$ is satisfied for any $x_m \in [0,1]$.

16. (a) Consider the explicit difference scheme

$$\frac{u_m^{n+1}-u_m^n}{\varDelta\tau} = a_m^n\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\varDelta x^2} + b_m^n\frac{u_{m+1}^n - u_{m-1}^n}{2\varDelta x} + c_m^n u_m^n$$

to the parabolic partial differential equation

$$\frac{\partial u}{\partial\tau} = a(x,\tau)\frac{\partial^2 u}{\partial x^2} + b(x,\tau)\frac{\partial u}{\partial x} + c(x,\tau)u.$$

Assume that its stability with respect to initial value and nonhomogeneous term is proved under certain conditions. Show that for its solution, under these conditions there is the following relation: $u\left(x,\tau;\Delta x,\Delta\tau\right)=u(x,\tau)+a\left(x,\tau;\dfrac{\Delta x^2}{\Delta\tau}\right)\Delta\tau+O(\Delta\tau^2)$, where $\left|O(\Delta\tau^2)\right|\le c\Delta\tau^2$, c being bounded as $\Delta\tau\to 0$ with $\dfrac{\Delta x^2}{\Delta\tau}=\mathit{constant}.$

(b) Suppose we have two such approximate solutions $u\left(x,\tau;\Delta x,\Delta\tau\right)$ and $u\left(x,\tau;\Delta x/2,\Delta\tau/4\right)$. Find a linear combination

$$(1-d)\times u\left(x,\tau;\Delta x,\Delta\tau\right)+d\times u\left(x,\tau;\Delta x/2,\Delta\tau/4\right)$$

such that it is an approximate solution with an error of $O(\Delta\tau^2)$.

17. (a) Assume that an approximate solution $u\left(x,\tau;\Delta x,\Delta\tau\right)$ has the following expression:

$$\begin{aligned}&u\left(x,\tau;\Delta x,\Delta\tau\right)\\&=u\left(x,\tau\right)+a\left(x,\tau;\frac{\Delta x}{\Delta\tau}\right)\Delta\tau^2+b\left(x,\tau;\frac{\Delta x}{\Delta\tau}\right)\Delta\tau^3+O\left(\Delta\tau^4\right),\end{aligned}$$

where $u\left(x,\tau\right)$ is the exact solution. Suppose that we have two approximate solutions: $u\left(x,\tau;\dfrac{1}{12},\dfrac{T}{16}\right)$ and $u\left(x,\tau;\dfrac{1}{9},\dfrac{T}{12}\right)$. Find a linear combination

$$(1-d)\times u\left(x,\tau;\frac{1}{12},\frac{T}{16}\right)+d\times u\left(x,\tau;\frac{1}{9},\frac{T}{12}\right)$$

such that it is an approximate solution with an error of $O(\Delta\tau^3)$.

(b) Suppose that there is another approximate solution $u\left(x,\tau;\dfrac{1}{15},\dfrac{T}{20}\right)$. Find a linear combination

$$d_0\times u\left(x,\tau;\frac{1}{15},\frac{T}{20}\right)+d_1\times u\left(x,\tau;\frac{1}{12},\frac{T}{16}\right)+d_2\times u\left(x,\tau;\frac{1}{9},\frac{T}{12}\right)$$

such that it is an approximate solution with an error of $O(\Delta\tau^4)$, where $d_0=1-d_1-d_2$.

18. *Explain why, how and when the extrapolation technique will improve the accuracy of numerical solutions.

19. Let $\mathcal{V}=\{u\,|\,u=(u_0,u_1,\cdots,u_{M-1},u_M)\}$ be the grid function space on $\Omega_h=\{x_m\mid x_m=x_l+mh,0\le m\le M,h=(x_u-x_l)/M\}$. For any $u\in\mathcal{V}$, and $v\in\mathcal{V}$, introduce the inner product

$$(u,v)=h\left(\frac{1}{2}u_0v_0+\sum_{m=1}^{M-1}u_mv_m+\frac{1}{2}u_Mv_M\right)$$

and norm

$$\|u\| = \sqrt{(u,u)}.$$

In addition, denote

$$\varDelta_x u_m = \frac{1}{2h}(u_{m+1} - u_{m-1}), \quad \delta_x^2 u_m = \frac{1}{h^2}(u_{m+1} - 2u_m + u_{m-1}).$$

(a) Suppose

$$a(x) \in C^{(2)}[x_l, x_u], \quad a(x) \geq 0, \quad a(x_l) = a(x_u) = a'(x_l) = a'(x_u) = 0$$

and

$$\max_{x_l \leq x \leq x_u} |a''(x)| = c_1.$$

Prove

$$\left(a\delta_x^2 u, u\right) \leq \frac{1}{2}c_1\|u\|^2.$$

(b) Suppose

$$b(x) \in C^{(1)}[x_l, x_u], \qquad b(x_l) = b(x_u) = 0, \qquad \max_{x_l \leq x \leq x_u} |b'(x)| = c_2.$$

Prove

$$(b\varDelta_x u, u) \leq \frac{1}{2}c_2\|u\|^2.$$

20. Suppose that $(a_{12})_{0n} = (a_{12})_{Mn} = (a_{12})_{m0} = (a_{12})_{mN} = 0$. Show

$$\begin{aligned}
& h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m-\frac{1}{2},n+\frac{1}{2}}\ u_{mn} \\
= & -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x u_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
& -h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}\left(\delta_x a_{12}\right)_{m+\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}}\ u_{m+1,n}
\end{aligned}$$

and

$$\begin{aligned}
& h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x\delta_y u_{m+\frac{1}{2},n+\frac{1}{2}}\ u_{mn} \\
= & -h_1 h_2 \sum_{m=1}^{M-1}\sum_{n=1}^{N-1}(a_{12})_{mn}\ \delta_x u_{m-\frac{1}{2},n}\ \delta_y u_{m,n+\frac{1}{2}} \\
& -h_1 h_2 \sum_{m=0}^{M-1}\sum_{n=1}^{N-1}\left(\delta_x a_{12}\right)_{m+\frac{1}{2},n}\ \delta_y u_{m+1,n+\frac{1}{2}}\ u_{mn}
\end{aligned}$$

by a direct calculation.

21. Suppose $\{u_m^k\}$ is the solution of the difference scheme

$$\begin{aligned}\frac{1}{\varDelta\tau}(u_m^{k+1}-u_m^k) &= a(x_m)\delta_x^2 u_m^{k+\frac{1}{2}} + b(x_m)\varDelta_x u_m^{k+\frac{1}{2}} + c(x_m)u_m^{k+\frac{1}{2}}\\ &\quad + g(x_m,\tau^{k+\frac{1}{2}}), \quad 0\le m\le M, \quad 0\le k\le K-1,\\ u_m^0 &= f(x_m), \quad 0\le m\le M,\end{aligned}$$

where $u_m^{k+\frac{1}{2}} = \frac{1}{2}\left(u_m^k + u_m^{k+1}\right)$ and

$$\begin{aligned}&a(x)\in C^{(2)}[x_l,x_u], \qquad b(x)\in C^{(1)}[x_l,x_u],\\ &a(x)\ge 0, \quad a(x_l)=a(x_u)=a'(x_l)=a'(x_u)=b(x_l)=b(x_u)=0,\\ &\max_{x_l\le x\le x_u}|a''(x)|=c_1, \quad \max_{x_l\le x\le x_u}|b'(x)|=c_2, \quad \max_{x_l\le x\le x_u}|c(x)|=c_3,\\ &c=c_1+c_2+2c_3, \qquad \varDelta\tau\le 2/[3(c+1)].\end{aligned}$$

Prove

$$\|u^{k+1}\|^2 \le \mathrm{e}^{3(c+1)T/2}\left(\|f\|^2 + \frac{3}{2}\varDelta t\sum_{l=0}^{k}\|g^{l+\frac{1}{2}}\|^2\right), \quad 0\le k\le K-1.$$